

BIRLA CENTRAL LIBRARY

PILANI (RAJASTHAN)

Call No

530

P41C

Accession No

11251

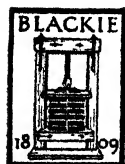
A COURSE OF PHYSICS

COURSE OF PHYSICS

BY

HENRY A. PERKINS, Sc.D.

Professor of Physics, Trinity College, Hartford



BLACKIE & SON LIMITED
LONDON AND GLASGOW

COPYRIGHT, 1938, BY
PRENTICE-HALL, INC.
70 FIFTH AVENUE, NEW YORK

ALL RIGHTS RESERVED. NO PART OF THIS BOOK MAY BE
REPRODUCED IN ANY FORM, BY MIMEOGRAPH OR ANY
OTHER MEANS, WITHOUT PERMISSION IN WRITING FROM
THE PUBLISHERS

First Printing.....March, 1938
Second Printing.....August, 1938
Third Printing.January, 1940

PRINTED IN THE UNITED STATES OF AMERICA

Preface

THE purpose of this book is to give the student a substantial grasp of physical principles rather than to describe phenomena. The more difficult portions are therefore treated more fully than is usual in elementary texts. Difficulties are met without evasion and without sacrificing clearness in an attempt to be brief. It is a great mistake to mask what is really hard by making it seem easy, and students are invariably confused by such a device. It is better to meet difficulties fairly and squarely, even if the text is lengthened by such a policy.

The aim has been to make explanations so clear that the student should be able to understand them without assistance. Such an aim seems rather obvious, yet all teachers of physics are familiar with the necessity of explaining an explanation and so wasting precious time. The lecturer should have his hour free for enlarging on the text, for citing applications and illustrations from everyday life, and for performing experiments.

The language of mathematics has been freely used even when a purely verbal discussion would be possible, because most students find a concise symbolic statement easier to grasp than a verbal one. However, only very simple algebra and trigonometry are used. The proofs of classical equations and theorems are in the main traditional proofs. There is no advantage in straining after new ways of deriving old relations unless there is a distinct gain in brevity or clarity. This is rarely possible, though here in a few instances new and simplified demonstrations have been introduced, but not for the sake of novelty.

The ground covered in this book is a little more extended than usual. The object of the broader program is to make the change from elementary physics to intermediate courses less violent than would otherwise be the case. But this has not made the book unduly long, and no fine print (except in the problems) has been used to make it appear shorter than it really is. Holding the extent of the text down to reasonable dimensions without sacrificing either clarity or scope has been made possible in two ways: The portions that any student grasps at once have not been dwelt on at great length, and very little space has been devoted to directions telling the instructor how to perform demonstration experiments in the classroom. He should know how to do them unaided.

Although modern ideas have been as fully treated as is feasible in

an elementary course, a certain amount of material whose importance is largely historical has been retained. This is partly because of its intrinsic interest and partly because of its pedagogical value in making clear fundamental concepts and paving the way for more difficult modern ideas. In electricity, the notion of electrons and protons is introduced almost at once, but the fuller discussion of atomic structure and allied topics is postponed until classical electromagnetism has been thoroughly treated. In this earlier section, electrons are referred to only when they are a genuine help to understanding what happens. There is no advantage in constantly referring to an electric current as a flow of electrons in the opposite direction, or in describing a positive charge as a deficit of electrons, after the idea has once been grasped. In fact, there is a very real danger that a too facile treatment of such delicate subjects as metallic conduction will give the student a half-baked or even false idea of recent physical thought.

The determination to be up-to-date at any cost can be very detrimental to an otherwise excellent textbook. This is because many recent discoveries have to be treated in the popular manner of the daily press. The reader is awed, but gains only a very hazy notion of reasons and causes. Physics which only describes and does not explain is not physics. The student should realize from the start this important distinction between an exact science and one which is concerned mainly with phenomena.

To sum up: the chief aim in writing this book has been to explain the difficult portions of physics fully and clearly, to introduce modern physical ideas wherever they can be discussed with some degree of rigor, and to retain those aspects of classical physics which are still valuable as the basis of its fundamental principles.

The author gratefully acknowledges much valuable assistance in the preparation of this book. Portions have been read and criticized by specialists in particular fields, and their advice has been most helpful. He wishes to express his sincere thanks to all of them, and in particular to his colleague, Professor Arthur P. R. Wadlund, who has rendered invaluable assistance both in reading the manuscript and in checking the solutions of most of the problems. Dr. Howard D. Doolittle, another colleague, has also read most of the manuscript, and deserves the author's sincere thanks for many helpful suggestions.

Finally, to Doctor Edward U. Condon belongs the credit for important improvements in the text submitted to him. His sound judgment and critical acumen as science editor for the publishers are greatly appreciated.

H. A. P.

Contents

PART I

MECHANICS

CHAPTER	PAGE
1. INTRODUCTORY	3
2. STATICS	17
3. KINETICS	33
4. GRAVITATION AND FALLING BODIES	44
5. WORK, ENERGY, POWER, AND FRICTION	57
6. MOTION IN A CIRCLE	79
7. ROTATION OF A BODY	93
8. ELASTICITY	109
9. HYDROSTATICS	116
10. MECHANICS OF GASES	130
11. FLUIDS IN MOTION	138
12. SURFACE TENSION AND CAPILLARITY	145

PART II

HEAT

13. TEMPERATURE	157
14. THERMAL EXPANSION	163
15. IDEAL GASES	172
16. HEAT MEASUREMENTS	185
17. CHANGE OF STATE	194
18. VAPORS AND GASES	206
19. RELATIONS BETWEEN THE STATES	214
20. HEAT AND ENERGY	221
21. SOLUTIONS	239
22. PROPAGATION OF HEAT	250

PART III

WAVE MOTION AND SOUND

CHAPTER	PAGE
23. WAVES	275
24. SOUND AND ITS TRANSMISSION	298
25. PROPERTIES OF SOUND	307
26. HEARING AND ACOUSTICS	320
27. THE PHYSICAL BASIS OF MUSIC	328
28. THE PRODUCTION OF TONES—VIBRATING SOLIDS	337
29. THE PRODUCTION OF TONES—VIBRATING GASES.	350

PART IV

LIGHT

30. PRODUCTION, PROPAGATION, AND PERCEPTION	363
31. REFLECTION	378
32. REFRACTION AT A PLANE SURFACE.	392
33. LENSES	397
34. OPTICAL INSTRUMENTS	417
35. DISPERSION AND SPECTRA	434
36. INTERFERENCE OF LIGHT	452
37. DIFFRACTION	463
38. FRAUNHOFER DIFFRACTION	473
39. POLARIZED LIGHT	486
40. COLOR	506
41. SOURCES OF LIGHT	516
42. OPTICAL PHENOMENA IN NATURE	531

PART V

ELECTRICITY AND MAGNETISM

43. MAGNETISM	541
44. ELECTROSTATICS	558
45. ELECTROSTATICS (<i>Continued</i>)	571
46. ELECTRODYNAMICS	589
47. THE ELECTRIC CURRENT	607
48. BATTERIES	626

CONTENTS

ix

CHAPTER	PAGE
49. THERMOELECTRICITY	637
50. ELECTRICAL MEASUREMENTS.	646
51. ELECTROMAGNETISM	661
52. INDUCED CURRENTS	676
53. ELECTRICAL MACHINERY	694
54. ELECTRICAL OSCILLATIONS	715

PART VI

CORPUSCULAR PHYSICS

55. ELECTRICAL DISCHARGES	725
56. THERMO- AND PHOTOELECTRIC EMISSION	741
57. X-RAYS AND RELATED PHENOMENA	754
58. ATOMIC STRUCTURE	767
59. RADIOACTIVITY	790

APPENDIX

THE SOLUTION OF PROBLEMS	807
INDEX	809
CONDENSED TABLE OF NATURAL TRIGONOMETRIC FUNCTIONS	821

PART I
MECHANICS

CHAPTER 1

Introductory

1. The scope of physics. The science of physics was formerly called *natural philosophy*. This meant that its purpose was to explain nature rather than merely to describe. In order to "explain" something we must show how it depends upon other things that we accept as true. These fundamental "truths" are called laws and are usually expressed in exact mathematical language. Their application to natural phenomena is worked out by mathematical analysis.

Nature in general is too complicated to be accounted for in the exact way just described, and so most of the things we see about us, trees, birds, and so forth, lie outside the realm of physics. But certain aspects of nature as well as man-made machines are sufficiently simple to be explained mathematically in terms of fundamental "laws." These aspects are the legitimate field of physics and are grouped for convenience under five heads: mechanics, heat, sound, light, and electricity. They are not wholly distinct from each other, but each deals with certain characteristic phenomena that give rise to problems capable of exact solution. In short, physics deals with the general principles and methods by which such problems are to be approached and solved.

2. Mechanics defined. Mechanics is the most fundamental of the five departments of physics. It treats of such ideas as motion, force, and energy and their relations to each other and to matter. But in general it is not concerned with different kinds of matter or its various properties such as color, temperature, or electrical condition. These questions are left mainly to other divisions of physics.

Mechanics may be divided, somewhat artificially, into two sections: **kinematics**, which treats of pure motion regardless of what causes it, and **dynamics**, the science of forces. Dynamics in turn has two divisions: **statics** and **kinetics**. The former treats of systems acted on by forces with no resulting motion, the latter of systems acted on by forces which result in motion.

3. Materials. As the materials of arithmetic and geometry are number and space, so the materials of mechanics may be said to be

time, distance, and mass. Time and distance are basic concepts which mean something very real to us but which we cannot define. Mass is not so directly perceived, but it can be defined in terms of force which is readily appreciated. We shall therefore take time and distance, and, for the present at least, force for granted. The meaning of mass will be explained farther on.

4. Measurements. The magnitude of a given time interval, or of distance or mass, can be measured only by comparing it with an accepted standard unit of the same kind. Thus, when we say that a man is six feet tall, we mean that a foot rule applied six times, so that at each application the new and old positions just touch, will reach from the floor to the top of his head. His height, then, is to the length of the ruler as six is to one, and we are really obtaining a ratio between two quantities of the same kind. In measuring time we use a similar process. The intervals between the ticks of a seconds pendulum are, so to speak, added end to end, and a time interval between two given events is described as containing so many elapsed seconds. So with mass, when we say that an object has a mass of twenty pounds we mean that twenty of our standard units of mass taken together have the same total mass as the object considered.

5. Standard units. There are two systems of units in common use in America and England. They are the yard, pound, and second of the British system, and the meter, kilogram, and second of the metric system. The latter is much simpler, is more scientific and has been legally adopted by all but a few backward nations, and the two just named.

The **second**, which is common to both systems, is defined as $\frac{1}{86,400}$ of a mean solar day, or the average elapsed time between the sun's successive crossings of a given meridian. This time differs slightly from day to day, but the average taken over a year is regarded as constant.

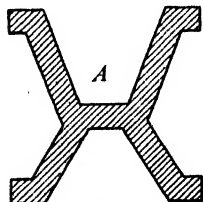


Fig. 1.

The **yard** and **pound** are arbitrary units based on the prototypes preserved in the Standards Office, Westminster, London, though in the United States the legal yard has been defined by Congress as $\frac{3}{4}$ meter.

The **meter** is defined as the distance between two fine lines ruled on a certain bar of platinum-iridium at the temperature of melting ice. This bar has a peculiar section as indicated in Fig. 1, and is thus very rigid, while the lines referred to are ruled on the surface A, so as to be in a plane along the bar's axis.

The meter was intended to be one ten-millionth part of the distance from the equator to the pole taken along the meridian of Paris. But the survey was not so accurate as was anticipated, so that the meter of the Archives, as it is called, is really arbitrary like the yard. It is kept at Sèvres, near Paris.

The **kilogram** is the mass of a piece of platinum also kept at Sèvres. It is almost exactly the mass of a liter of water at the temperature of 4°C , when it is most dense. Actually the standard kilogram weighs 0.04 gram more than was intended. But for ordinary calculations, it is sufficiently accurate to assume that a liter of water at 4°C weighs a kilogram, and that a cubic centimeter weighs a gram.

6. The c.g.s. system. A system based on the centimeter, gram, and second is used almost exclusively by physicists. They have substituted the gram (one thousandth of a kilogram) and the centimeter (one hundredth of a meter) for the kilogram and meter as being more suitable for measuring the magnitudes usually met with in the laboratory. All the other quantities of mechanics such as velocity, force, and so on, are expressible in terms of the c.g.s. system. With the additional notions of temperature, and possibly magnetic permeability and the dielectric constant (to be defined later), all the quantities used in physics may be derived from these basic quantities.

7. Compound quantities. Such quantities as velocity, volume, density, and so forth, are compounds of the basic quantities. Velocity, for instance, is measured in terms of distance divided by time, as is evident from the term miles *per* hour, feet *per* second. The preposition *per* means of course *through* or *by*. Our only true speed unit, the *knot*, is a nautical mile per hour, and as $69\frac{1}{2}$ statute miles equal 60 nautical miles, the knot is about 15.3 per cent faster than a mile per hour.

It should be noted that this very old unit of speed is named from the method used in "heaving the log" to determine a vessel's motion through the water. The log line, knotted at regular intervals, slipped through the fingers, and the number of knots felt in a given time measured the speed. Therefore a knot is *not the distance* of a nautical mile, as is frequently stated, and the expression "ten knots an hour" is incorrect, "ten knots" being a complete statement of the facts.

Other compound quantities are area (the square of a length), volume (length cubed), and density, which is the amount of matter per unit volume of a substance. Density is measured in terms of mass per unit length cubed.

8. Dimensions. The way in which the fundamental quantities, mass, distance, and time, enter into a compound quantity gives rise to what are called "dimensional" formulae. These "dimensions" are expressed in terms of L , M , and T , the initials of the three basic quantities. Thus the dimensions of velocity are obviously L/T , or $[LT^{-1}]$ as it is usually written, the bracket indicating that we are dealing with a dimensional expression, and not an ordinary algebraic equation. Setting v for velocity, we would then obtain the complete expression $[v] = [LT^{-1}]$. Similarly, if A represents area, and if V represents volume, $[A] = [L^2]$ and $[V] = [L^3]$. Density may be expressed by $[d] = [ML^{-3}]$, which is mass divided by volume.

This method of expressing the essence, as it were, of a compound quantity is of the utmost value, often helping to formulate laws, and always acting as a check on equations, because the two members must of course have the same dimensions. We cannot equate area with volume, or velocity with density!

9. Angles. Everyone is familiar with the measurement of angles in terms of degrees, minutes, and seconds, but these units are wholly arbitrary and based on no physical or mathematical considerations whatever. The circle might just as well, perhaps better, have been divided into four hundred degrees, for instance.

The logical angular unit, used in physics, is the *radian*, but before we can appreciate its significance, it is necessary to explain the mean-

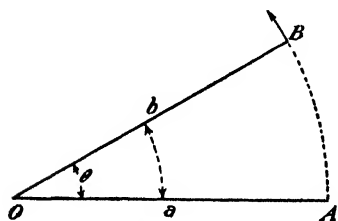


Fig. 2.

ing of angles. Consider the line OB in Fig. 2 as having started in coincidence with OA and to have developed the angle θ by turning about O in a counterclockwise sense, as indicated by the arrow. This sense of rotation is considered positive, and the angle θ is a positive angle. If OB had been rotated in a clockwise sense from the

horizontal, it would have developed the negative angle $-\theta$. The magnitude of the angle depends upon the portion of a revolution executed by the rotating radius OB and is independent of the length of that line.

10. Measurement of angles. The absolute magnitude of any angle, such as θ in Fig. 2, is the ratio of the arc BA to the length of the radius of that arc, or OA . This is known as **circular measure**. It is the only natural and logical measure because it has no arbitrary unit, such as the degree, and is independent of the particular arc or radius

chosen. This may be seen by taking another arc, as ba ; then, by geometry, the ratio of ba to its radius Oa equals BA/OA , and both ratios measure the same angle.

When the length of the arc is equal to its radius, the ratio is unity and the angle thus determined is the natural unit of circular measure. It is called the **radian** and is equal to $57.296+$ degrees. This is calculated by means of the consideration that its arc, which has the same length as its radius, by hypothesis, goes 2π times into the circumference. There are therefore 2π radians in a circle. But a circle contains 360 degrees; therefore one radian has $360/2\pi$ degrees, which is the value given above. It also follows that $360^\circ = 2\pi$ radians, $180^\circ = \pi$ radians, and $90^\circ = \pi/2$ radians.

11. Curvature. A circle has constant curvature, denoted by σ , and this is defined as the change in direction per unit length of arc. The direction of a curve at any point is that of its tangent at that point. Therefore in a circle the curvature indicated in Fig. 3 is given by

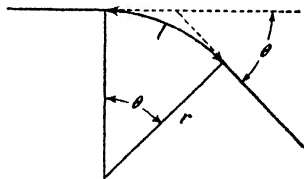


Fig. 3.

$$\sigma = \frac{\theta}{l}. \quad (1)$$

But in circular measure

$$\theta = \frac{l}{r}. \quad (2)$$

Substituting the value of θ from (2) in (1) we obtain

$$\sigma = \frac{1}{r}. \quad (3)$$

This means that curvature is numerically equal to the reciprocal of the "radius of curvature."

For all other curves, the curvature at any point is that of the circle which most nearly fits the curve at that point. Thus in Fig. 4, op and OP are the radii of curvature at the points p and P , and the curvatures are $1/op$ and $1/OP$ respectively. Since curvature increases as its radius decreases, a pin point has a very large curvature, while that of the surface of the earth is extremely small.

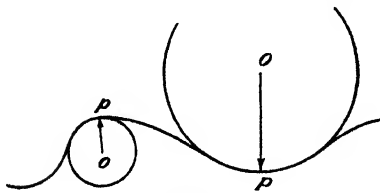


Fig. 4.

12. Motion. This important physical concept may be defined as progressive change of position of a point or a body during an interval of time. The time element is essential, for if a body assumes two successive positions *one* must have followed the other *in time*. When stress is laid on the time consideration, motion is thought of as a speed or velocity, but if we are interested only in the changing positions regardless of how long a time was required, we concentrate our attention on the *path* followed, which is the locus of succeeding positions of the moving body.

A *particle* is a body whose dimensions may be neglected as having no significance in the problem under consideration. A *rigid body* is one whose dimensions are significant, and it may be regarded as made up of particles at fixed distances from each other.

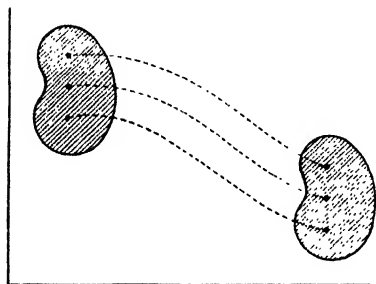


Fig. 5.

The path of a moving *particle* is a straight or curved line. The path of a moving *body* is the totality of the paths of its component particles. If these paths all have the same length as shown in Fig. 5, the motion is pure translation. Translation may take place along either a straight or a

curved path. At the end of the displacement, or any portion of it, the body is oriented in the same way with reference to any fixed plane.

A body is said to rotate when every particle in it describes concentric circles around some common axis as in Figs. 6 (a) and 6 (b). In (a) the axis lies outside the body, in (b) it passes through it, but both are rotations because the component particles all move in concentric circles. Translation and rotation may take place at the same time. In that case the body at any instant is rotating about an instantaneous center which itself is moving along a straight or curved path. The simplest case is that of a wheel rolling along level ground. It rotates about its axis which at the same time is moving horizontally in a straight line.

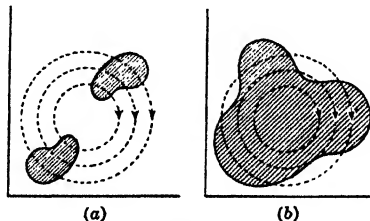


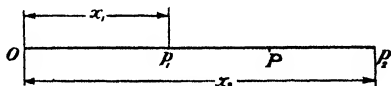
Fig. 6.

Any motion, however complicated it may appear, can be resolved

into a pure rotation about an instantaneous axis, as in Fig. 6, and a pure translation of the body as a whole, as in Fig. 5.

13. Linear velocity. When a particle moves along a straight line, the motion is rectilinear and its velocity is defined as the time rate of change of position measured along the line.

If the velocity is constant it can be calculated by dividing the distance between any two instantaneous positions by the time required to pass from one to the other. Thus let x_1 and x_2 in Fig. 7 represent the distances of the positions p_1 and p_2 from a point O on the path. Then



$x_2 - x_1$ is the distance traveled, and if the time at p_1 is t_1 and at p_2 is t_2 , then

$$v = \frac{x_2 - x_1}{t_2 - t_1}. \quad (1)$$

If the velocity is not constant, but varies in any conceivable manner, equation (1) gives us the *average* velocity over the time interval $t_2 - t_1$. If we then wish to know the velocity at some particular point P , we find it by bringing p_1 and p_2 very close together and at either side of that point. In this way $x_2 - x_1$ and $t_2 - t_1$ become vanishingly small and the instantaneous value of the variable velocity is expressed in the notation of the calculus by $v = dx/dt$. This is the limiting value of the expression (1) above as p_1 and p_2 approach the desired point. The term dx means an infinitesimal distance corresponding to the elapsed time dt ; therefore any velocity, constant or changing in the x direction, may be indicated by dx/dt .

14. Acceleration. The time rate at which a body gains or loses velocity is called **acceleration**. It is expressed in terms of *unit displacement per second each second*, and its more usual units are centimeters, or feet, per second per second.

When acceleration is constant, we may calculate it by dividing the gain in velocity by the time required to make this gain or

$$a = \frac{v_2 - v_1}{t_2 - t_1} = \frac{v_2 - v_1}{t}, \quad (1)$$

where v_1 and v_2 , t_1 and t_2 are the initial and final velocities and times, and t is the time interval. If, for instance, an automobile going at 20 feet per second speeds up in 2 seconds to 30 feet per second at a uniform rate of increase, then its acceleration is at the rate of 5 feet per second each second, or 5 ft./sec².

But when acceleration is variable, so that the body gains velocity at different rates during the time t , the above statement is not true, and we must express a by the notation of the calculus as in the case of variable velocity. We then write $a = dv/dt$, which is the limiting value of the average acceleration when v_2 and v_1 approach each other as the time interval $t_2 - t_1$ approaches zero.

Equation (1) may be transposed to read

$$at = v_2 - v_1. \quad (2)$$

In this form it enables us to calculate the velocity acquired in a given time under a given constant acceleration, and thus it may be written simply $v = at$, where v is the velocity gained during the time t . As the dimensions of velocity are $[LT^{-1}]$, the dimensions of acceleration are those of velocity divided by time, or $[LT^{-2}]$.

15. Vectors and scalars. Such magnitudes as displacement, velocity, and acceleration are usually associated in our minds with the idea of direction, while the ideas of mass and time do not involve direction. When a quantity has both direction and magnitude, it is called a **vector** quantity. When it has only magnitude it is called a **scalar** quantity. The magnitude of a vector quantity considered independently of direction is itself a scalar quantity. Thus velocity is essentially vectorial, but we may be interested only in the *rate*, as in discussing a sprinter's record, regardless of the direction of the race. This is a scalar quantity designated by *speed*. Similarly *length* may be used to designate the scalar aspect of the vector quantity *displacement*.

A symbol used to designate a vector quantity is referred to as a *vector* without the addition of the word quantity. Thus vector quantities may be expressed in terms of vectors, just as scalar quantities may be expressed by numbers. In a diagram a vector is represented by an arrow pointing in the proper direction, and having a length proportional to the vector's scalar magnitude. The point of the arrow is called the *terminus* and the "nock" is the *origin* of the vector. In analytical expressions it may be designated by a line above the letter or letters indicating the vector, as \overline{AB} , or by printing the symbols in bold-faced type, as \mathbf{R} .

16. Addition of vectors. Scalar quantities of the same kind are combined by algebraic addition or subtraction, and due account should be taken of their signs, either positive or negative. Vectors, on the other hand, can be added or subtracted only geometrically; this property is a third criterion of a true vector. It must then have

magnitude, direction, and be capable of geometrical (or vector) addition and subtraction with other vectors of the same kind.

Vector addition is best shown by the so-called parallelogram of vectors. Let A and B be two vectors, as in Fig. 8, having scalar magnitudes of 5 and 10 units, and directed northeast and east respectively. Their vector sum R is obviously not 15 units, as if they were scalars, but is less than that, and its direction lies somewhere between their directions. To

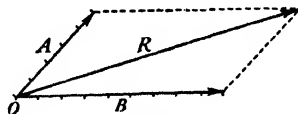


Fig. 8.

prove that the diagonal of the parallelogram correctly represents the sum, we may reason as follows: It is a principle of geometry that two directions perpendicular to each other are mutually independent. If we travel due north, for instance, we get no farther east or west; therefore, we may consider A as made up of two independent components indicated in Fig. 9. One of these goes north, the other east, and the numerical value of each is $5 \cos 45^\circ = 3.54$. The northerly component has no effect upon the easterly vector B , but A 's easterly component b' is added arithmetically to B , which gives a total "easting" of 13.54, as shown in Fig. 10. Similarly the northerly component, a' of A produces a "northing" of 3.54 that is unaffected by the two easterly vectors, and the

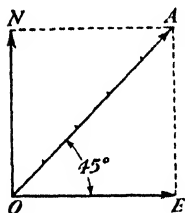


Fig. 9.

sum, or *resultant*, is drawn to the point specified by 13.54 units east and 3.54 units north of the origin, as shown in Fig. 10. If A' represents the side of the parallelogram opposite A , since it is equal in direction and magnitude, it represents the same vector quantity. Therefore we might have saved the construction line B' by drawing only B and A' , placing the origin of A' at the terminus of B and then uniting the origin of B with the terminus of A' to form R . R could also have been formed by uniting B' to A in a similar manner. This addition by a triangle rather than a parallelogram is usually the better way, but it should always be remembered that it involves origin to terminus contact, while the parallelogram requires the vectors to be placed origin to origin.

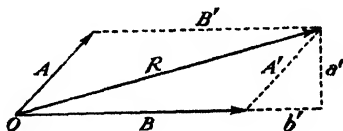


Fig. 10.

17. Subtraction of vectors. A vector difference is obtained by subtracting one vector from another *vectorially*. This is accomplished

by reversing the vector to be subtracted (subtrahend) and then obtaining the resultant as above, just as in algebra the difference $a - b$ may be thought of as the sum $a + (-b)$. Thus if we wish to subtract vector A in Fig. 10 from B , A is reversed and the parallelogram of Fig. 11 results.

Vector differences are useful in finding relative magnitudes and directions; for instance, in comparing two velocities. If an airplane,

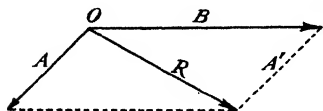


Fig. 11.

moving with the wind, is going at a velocity of 60 miles per hour over the ground and the wind has a velocity of 20 miles per hour, then the velocity of the airplane through the air, or its motion *relative* to the wind, is 40 miles

per hour. But if we wish to know the velocity of the air with reference to the plane, or *apparent* wind velocity as observed by the aviator, we must subtract the plane's velocity from that of the wind, obtaining -40 miles per hour. The minus sign indicates that the wind appears to blow against the plane, although it really blows with it. Therefore to obtain relative motion, subtract the velocity of the object with reference to which the relative motion is desired.

The rule just stated is equally true when the vectors are not in the same straight line. For instance, if it is desired to find the motion of a west wind blowing 30 miles per hour with reference to an airplane flying north with the same velocity, we reverse the vector representing the plane's velocity and add vectorially to that of the wind. This gives us a wind velocity of $30\sqrt{2}$ miles per hour blowing apparently from the northwest. The aviator therefore encounters a stronger wind blowing at an angle of 45° from head on, instead of from the side.

18. Graphic summation of many vectors. When there are more than two vectors we may first add two of them either by the parallelogram or triangle methods, then add the third to the *resultant* thus

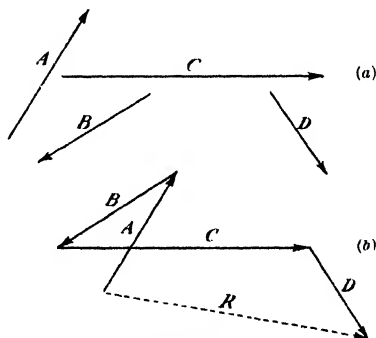


Fig. 12.

obtained, then the fourth to the sum of the first three and so on. But it is much simpler and better to proceed as indicated in Fig. 12. The vectors shown in (a) have been added in (b) by placing them head

The square of the resultant equals the sum of the squares of its x , y , and z components. This proposition may be demonstrated from Fig. 17 as follows: The vector Op is the hypotenuse of a right-angled triangle of which x and y are the other two sides. Therefore $\overline{Op}^2 = x^2 + y^2$. But Oq , or R , is the hypotenuse of the right-angled triangle Oqp , so $R^2 = \overline{Op}^2 + z^2$, and substituting for Op we obtain

$$R^2 = x^2 + y^2 + z^2, \quad (1)$$

or

$$R = \sqrt{x^2 + y^2 + z^2}.$$

SUPPLEMENTARY READING*

Emile Borel, *Space and Time*, Blackie & Son, London, 1926.

H. A. Erikson, *Elements of Mechanics* (Chap. 1), McGraw-Hill, 1927.

A. P. Wills, *Vector and Tensor Analysis* (First ten pages), Prentice-Hall, 1931.

PROBLEMS†

1. How many radians are subtended by a 12 ft. arc of a circle whose radius is 3 ft.? How many degrees does this represent? *Ans.* 4 radians, 229.2 degrees.

2. Convert 30° , 45° , 60° , and 90° to radians. *Ans.* 0.52; 0.79; 1.05; 1.57 radians.

3. What is the curvature of an arc 10 ft. long which subtends an angle of 4 radians? *Ans.* 0.4 ft^{-1} .

4. A train passes a signal tower at 12:30 P.M. and a switch three miles beyond at 12:35. How fast is it going if the speed is constant? *Ans.* 52.8 ft./sec.

5. A ship heads due east with a speed of 15 knots, across a southerly tidal current of 5 knots. What is its resultant velocity? What is the angle its course makes with the north-south meridian? *Ans.* 15.8 knots; $71^\circ 5'$ east of south.

6. A man who swims in still water at 2 mi./hr. crosses a stream half a mile wide by swimming straight across the current. He lands a quarter of a mile downstream. How swift is the current? How long is he in crossing? What is his actual speed? *Ans.* 1 mi./hr.; 15 min.; 2.24 mi./hr.

* A few of the books recommended at the end of this and following chapters are somewhat too difficult for the usual beginner. They are listed in the hope that a few ambitious students will at least turn over their pages in the college library and obtain some impression of what lies before them. Also the instructor may find some of them useful in refreshing his memory of topics outside his own special field.

† Before working any problems, the student should read the brief appendix entitled "The Solution of Problems."

7. If the same swimmer of Problem 6 heads upstream so as to go straight across and it takes him 20 minutes, how swift is the current? *Ans.* 1.32 mi./hr.

8. What is the horizontal velocity of a stone thrown with a speed of 200 ft./sec. at right angles from an airplane traveling at 150 ft./sec., if air resistance is not considered? *Ans.* 250 ft./sec.

9. If the above stone was thrown eastward at the airplane traveling 200 ft./sec. northward, with what velocity would it strike, and from what direction? *Ans.* 282 ft./sec. from the northwest.

10. Calculate the resultant, and its direction, of two vectors whose scalar magnitudes are 9 and 5, if their angular separation is 60° . *Ans.* 12.29; $20^\circ 6'$ from the larger vector.

11. Calculate the resultant of the following vectors: 8 ft. north; 6 ft., 30° east of north; 10 ft. southeast; 4 ft., 30° west of south, and 5 ft. west. *Ans.* 4.06 ft. +; 41° north of east.

12. A ship sails 30° north of east at a speed of 20 knots. What are its component velocities northeast and east? *Ans.* 7.3 knots east; 14.1 northeast.

must be parallel or come together at a common point. They are then said to be *concurrent*. They must also lie in the same plane as was assumed above. This condition is described as *coplanar*. Finally, their vector sum must be zero.

In all such problems we must first find the point P where the forces meet. This point is then regarded as independent of its surroundings. Such an assumption is perfectly reasonable, for consider a single rivet in the steel frame of a skyscraper. It is at rest, and therefore whatever forces act upon it must balance each other regardless of all the rest of the building.

Having selected the appropriate point, we then decide which forces act upon it, and as we are now considering only problems concerning three forces, we know that their vectors can be formed into a closed triangle. The final solution consists in solving the triangle. This is always possible when we know the length of at least one side and two other "elements," that is, two other sides, or one side and one angle, or two angles.

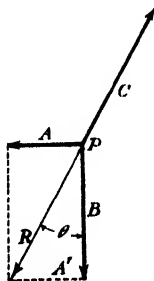


Fig. 20.

1. Suppose a boy in a swing is pulled sidewise by a horizontal force A until the ropes supporting him make the angle θ with the vertical, as shown in Fig. 20. The forces acting where he sits at P are his weight B , the pull A , and the tension C on the rope. As these are balanced, the resultant R of A and B is equal and opposite to C .

Therefore A , B , and C would form a right-angled triangle equivalent to $A'BR$. Suppose $\theta = 30^\circ$, and B is 100 lb.; then the force A is $100 \tan 30^\circ = 57.7$ lb., and the tension C is $100/\cos 30^\circ = 115.5$ lb.

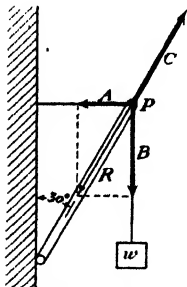


Fig. 21.

2. An almost identical problem is that of a bracket supporting a weight w as shown in Fig. 21. Here there is a compression of the diagonal strut whose reaction is shown by the vector C . This balances the resultant R of the forces A and B . Thus, as before, there is equilibrium between A , B , and C , and if θ is 30° and w is 100 lb., the tension in the horizontal member is 57.7 lb., and the compression in the strut is 115.5 lb.

3. If a weight is suspended from the center of a loose light cord supported at Q and S as shown in Fig. 22, the cord assumes the position QPS . The vector C , which represents the weight, is balanced by the resultant R of the two tensions A and B . Then C (or R) equals

$2A \sin \theta$. The angle θ may be found if the length $2l$ of the cord is known as well as the vertical height h between P and the line QS .

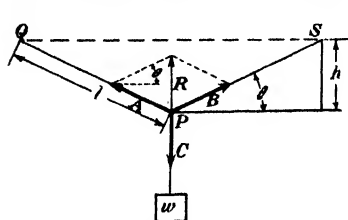


Fig. 22.

Then

$$\sin \theta = \frac{h}{l},$$

$$C = \frac{2Ah}{l},$$

and

$$A = \frac{Cl}{2h}.$$

If θ is less than 30° , l is greater than $2h$, and the tension A in the cord is greater than the force C . If h is very small compared to l , the tension is much greater than the weight causing it, and the cord may be broken by a surprisingly small load.

4. A roof truss, shown in Fig. 23, is similar to the bent cord. The force C exerted by the weight of the roof is supposed to be concentrated at P , where it is balanced by R , the resultant of the outward thrusts A and B of the beams PQ and PS . Then $R = 2A \sin \theta = C$. The lower ends of the beams rest upon the walls of the house, each exerting a downward force equal to half of C . This is balanced by the upward reactions D , but the components of C acting along the

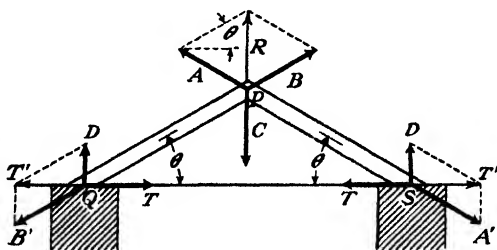


Fig. 23.

beams and represented by A' and B' tend to make their lower ends spread apart. Such action is neutralized by a tie rod QS which sustains a tension T equal and opposite to T' , the resultant of D and B' , or of D and A' . Therefore $T = \frac{1}{2}C/\tan \theta$.

5. Still another illustration of three balanced forces is found in a weight supported by a crane as shown in Fig. 24. Here the weight exerts a downward force A at the point P . This causes a tension B in the rope PQ . The resultant R of these two forces acting at P is

balanced by C , the outward thrust of the boom SP . Then

$$C \text{ (or } R) = \frac{A}{\tan \theta},$$

and $\tan \theta = \frac{h}{l}.$

$$\therefore C = \frac{Al}{h}.$$

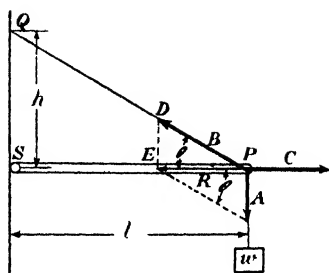


Fig. 24.

The tension in the rope is also given by $B = A/\sin \theta$ where $\sin \theta = h/\sqrt{h^2 + l^2}$, so that if h and l are known, both the compression of the boom and the tension in the rope due to w may be found.

25. Pressure. The word *pressure* has a very special meaning in physics. It does not mean just a pushing force in a vague sort of way. But it does mean the *intensity* of that force. Why is it that a pound push on a needle point pressing upon the skin hurts more than the same push acting on the dull point of a pencil? The common-sense answer is that one is sharper than the other. Physics is only highly developed common sense and it makes the same answer, but a more precise and useful one. It tells us that the prick of the needle is due to the fact that the force is concentrated over a very small area. The smaller the area the sharper is the prick when the same force is applied. Therefore pressure varies directly as the force and inversely as the area over which it is applied. It is *force per unit area*, or in symbols,

$$p = \frac{F}{A}. \quad (1)$$

This definition applies wherever the force is distributed evenly over an area normal to the direction of the force. The pressure on the bottom of a cubical tank containing water may be measured by the weight of the water in pounds divided by the area of the base given in square feet or square inches. Similarly the pressure of the atmosphere is due to its weight and is commonly stated in pounds per square inch.

26. Moment of a force. If you wish to open a heavy gate, where do you push? As far from the hinges as you can, of course. In what direction do you push? Naturally, at right angles to the gate. Everyone learns these two facts in early childhood, probably from "trial and error." At any rate, when you push in the proper way,

the force you exert is of greater account or *moment* than if you push near the hinges or at some other angle than ninety degrees. This use of the word *moment* is similar to its use in such expressions as "an event of great moment." It tells us that in making something rotate, the same force may have more or less *importance*, depending upon its direction and the point of application.

In order to measure this moment, or importance, of a force, consider Fig. 25 where a wheel pivoted at O is acted on by a force F which pushes normally against a crank r . The moment of the force is the product of F and r , for clearly the effectiveness of the push depends directly upon both quantities. If you increase F , keeping r constant, the turning moment increases, but if you use a correspondingly longer crank, you obtain the same result without changing the force. If the same force

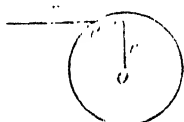


Fig. 25.

were applied at the point p instead of at the end of r , the moment would be exactly the same. What counts is not the point of application, but the *perpendicular distance between the axis and the line of action of the force*. So to obtain the moment of a force, drop a perpendicular from the axis upon the line of action of the force. This is the *lever arm*, and when multiplied by the force gives the moment in a compound unit such as pound-feet. The moment is said to be positive if it tends to produce a counterclockwise rotation. This is in accordance with the convention regarding positive and negative angles.

In Fig. 26, the above rule is further illustrated by a force F acting on a block B . Its line of action is produced to obtain the lever arm r by dropping a perpendicular from the axis at P .

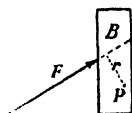


Fig. 26.

27. Couples. If a log of wood is lying on the ground, and if two men push against its ends equally hard, but in opposite directions at right angles to the log, it is easy to see what will happen. The log will spin around, about a center midway between the two men. It would be a very poor way to try to push the log along a road, for all one obtains is pure rotation. This arrangement of forces is shown in Fig. 27 and is called a *couple*. The forces must be equal, oppositely directed, parallel but not in the same line (nonconcurrent) and must lie in the same plane (coplanar). No rotation is ever produced except when a couple is acting. A force applied to the rim of a wheel turning freely on an

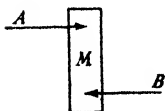


Fig. 27.

axle, makes the wheel turn, but only because an equal and opposite reaction holds the axle in place.

The value of a couple's moment is called its **torque**.† It is calculated by taking the *product of one of the two forces and the perpendicular distance between their lines of action*. Thus in Fig. 28, a body M is acted on by the couple AB . If P is taken as the axis or *center of moments*, the force A has a positive moment Aa , where a is its lever arm with respect to P . Similarly the moment of B is positive and equal to Bb . The total torque L is therefore $Aa + Bb$.

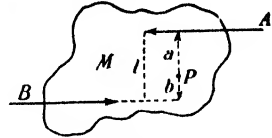


Fig. 28.

But $A = B$; hence $L = A(a + b)$, or $L = Al$ as stated above.

In the case just discussed the axis was supposed to lie between the two forces, but the value of the couple is the same no matter where P is placed. Take, for instance, the two forces shown in Fig. 29 with P outside their lines of action. Now the moment of A is $+Aa$, while that of B is $-Bb$ because B tends to produce a clockwise rotation. The net torque is then

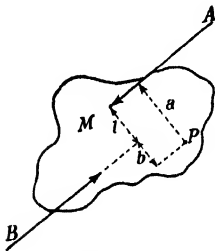


Fig. 29.

$$Aa - Bb = A(a - b) = Al$$

as before. From this it follows that the torque due to a couple is independent of its position in a given plane. We may move the figure representing a couple freely about the plane, and we may also transform

it by decreasing F and increasing l proportionately, and vice versa.

28. Equilibrium of bodies. The wheel of a moving automobile is being both rotated and translated. But if the brakes are set tight and the wheel skids, there is then only translation. The torque due to the brakes has neutralized the torque which caused rotation. On the other hand, attempting to start the auto on an icy pavement results only in spinning the wheels. Here is rotation without translation because the feeble tractive effort exerted on ice is neutralized by the car's resistance to starting.

From the preceding illustrations we may infer that to prevent a couple from producing rotation, it must be opposed by another couple, and to prevent a force from producing translation it must be opposed by another force. In general, if the *algebraic* sum of all the

† *Torque* is also used in speaking of a moment due to a single force, though this is not strictly correct.

turning moments acting on a body with reference to any axis, is zero, there is no rotation; and if the *vector* sum of all the forces is zero, there is no translation. These principles taken together are the necessary and sufficient conditions for equilibrium. They may be stated symbolically as

$$\Sigma F = 0 \text{ (no translation),} \quad (1)$$

and

$$\Sigma L = 0 \text{ (no rotation).} \quad (2)$$

If (1) is true, but not (2), pure rotation results. If (2) is true, but not (1), pure translation results. If neither holds, the motion which results is a combination of both translation and rotation.

29. Parallel forces. When the forces acting on a body are all parallel to each other, the calculation of their resultant and the torques they produce is very simple, especially if they are also coplanar. Suppose a body is acted on by five parallel coplanar forces,

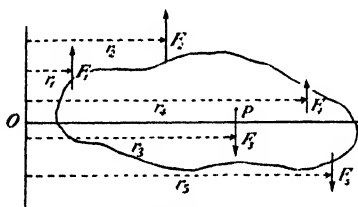


Fig. 30.

as shown in Fig. 30. Three acting upward are positive and two acting downward are negative. If there is no translation condition, (1) gives us

$$\Sigma F = F_1 + F_2 - F_3 + F_4 - F_5 = 0. \quad (1)$$

If there is no rotation, the algebraic sum of the torques around any

axis also vanishes. If we take any point as O for the center of moments, the various lever arms r may be found by dropping normals from the force vectors upon an axis parallel to these forces and passing through the center of moments. This fulfills the requirements stated in Article 26, that lever arms are the perpendicular distance between the axis and the line of action of the force. Then ΣL (or ΣFr) = 0. Taking account of the fact that three of the forces tend to produce positive rotation, and two negative rotation, we have

$$F_1 r_1 + F_2 r_2 - F_3 r_3 + F_4 r_4 - F_5 r_5 = 0. \quad (2)$$

In problems involving coplanar parallel forces in equilibrium, the unknown quantity is either one of the forces, or its point of application, or both. If the force is required, the force equation is used. If its point of application is required, the torque equation is used. If both are required, both equations are necessary.

Suppose that in Fig. 30, the force F_5 needed to prevent translation

is required. Let the other four forces have the values $+3$, $+2$, -4 , and $+2$ units respectively. Then equation (1) becomes $3 + F_5 = 0$, or $F_5 = -3$. It should be noted that as we do not know in advance the direction of F_5 , it must be given the positive sign in equation (1). Then if the result is negative we know that it acts downward.

If F_5 is known, but not its point of application, in order to prevent rotation, we must know the other lever arms as well as all the forces. Let the four known values of r be 1, 2, 3, and 4 units respectively. Then equation (2) becomes

$$3 \times 1 + 2 \times 2 - 4 \times 3 + 2 \times 4 - 3r_5 = 0,$$

which gives us $3r_5 = 3$, or $r_5 = 1$. This means that F_5 must act opposite to F_1 , instead of at the point shown in the diagram, in order to prevent rotation. We have thus found both the magnitude and location of a single force needed to balance the four given parallel forces so as to prevent both translation and rotation.

The choice of the center of moments is arbitrary. We could have taken P instead of O as such a point. This reduces the moment of F_3 to zero, and the lever arms measured to the left of P are negative. Equation (2) would then become

$$-3 \times 2 - 2 \times 1 - 4 \times 0 + 2 \times 1 - 3r_5 = 0,$$

or $3r_5 = -6$, and $r_5 = -2$, which places F_5 opposite to F_1 as before.

30. Three parallel forces. The special case of three forces occurs so often in simple structures and machines that it deserves particular mention. One such problem is represented in Fig. 31, where two weights are hung at the ends of a rigid bar of length l and whose own weight we will suppose negligible compared to w_1 and w_2 . It is desired to find the point P where it must be supported by a force F_3 in order not to rotate. Since

$$\Sigma F = 0,$$

$$-F_1 + F_3 - F_2 = 0.$$

$$\therefore F_3 = F_1 + F_2. \quad (1)$$

Taking moments around the left end of the bar, we find that

$$\Sigma L = 0 = -F_1 \times 0 + F_3 x - F_2 l. \quad (2)$$

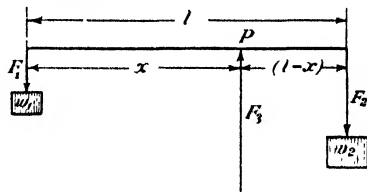


Fig. 31.

Substituting F_3 from (1) in (2) we obtain

$$(F_1 + F_2)x = F_2l.$$

$$\therefore \frac{l}{x} = \frac{F_1 + F_2}{F_2}. \quad (3)$$

Then by "division"

$$\frac{l-x}{x} = \frac{F_1}{F_2}. \quad (4)$$

Thus it appears that the point P divides l into two segments whose lengths are inversely proportional to the two weights. If the weight of the bar, however, must be considered, it appears as a third downward force acting at the bar's center of gravity, which is a point to be defined later. In this case the above simple solution is no longer valid.

Another problem of this type is that of a bar of length $2l$ and negligible weight supported at its ends by two men exerting forces

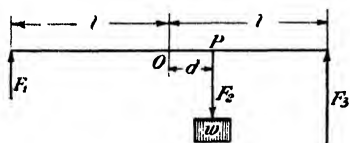


Fig. 32.

F_1 and F_3 , and carrying a weight w hanging from the point P on the bar. It might be required to find the position of P for any assigned distribution of the load between the two men, or to find F_1 and F_3 with P fixed. Let us consider the

latter case, taking our center of moments at the center of the bar. Then since $\Sigma F = 0$,

$$F_1 + F_3 = F_2,$$

and since

$$\Sigma L = 0,$$

$$F_3l = F_1l + F_2d.$$

Eliminating F_2 by substituting $F_1 + F_3$ for it, we obtain

$$F_3l = F_1l + F_1d + F_3d,$$

and

$$\frac{F_3}{F_1} = \frac{l+d}{l-d}, \quad (5)$$

which again shows that the two supporting forces are inversely proportional to the segments into which the bar is divided.

31. Center of gravity. Any body may be regarded as made up of a great number of very small particles, each having a finite mass. The forces of gravity acting upon these mass-particles are all practically parallel to each other unless the body is so large that the

gravitational forces directed toward the center of the earth have appreciably different directions. Let us assume, then, a body of moderate size, the mass-particles of which are acted upon by a system of parallel forces each proportional to the mass of the particle. *Then the center of gravity is a point through which the equilibrant of all the forces must act in order to produce equilibrium with the body in any position.* This means a point at which the body may be supported and placed in any position without having a tendency to rotate.

The center of gravity of a thin sheet may be found by experiment as follows: A piece of stiff cardboard cut in any shape is suspended by a thread from a point on its edge. It will hang with its center of gravity directly under the support, because only then does the sum of all the infinitesimal turning moments with reference to the support equal zero. Therefore the center of gravity lies on a line drawn on the cardboard vertically downward from the support. It is then suspended from another point on the edge and a similar vertical line is drawn. The intersection of these two lines is the point sought, and the cardboard should balance when supported there.

32. Calculation of the center of gravity. In Fig. 33 let the forces 2, 3, and 1 lb. act respectively on the mass particles m_1 , m_2 , and m_3 . Applying the rule $\Sigma F = 0$, we find that the equilibrant E must be 6 lb., equal and opposite to the resultant R .

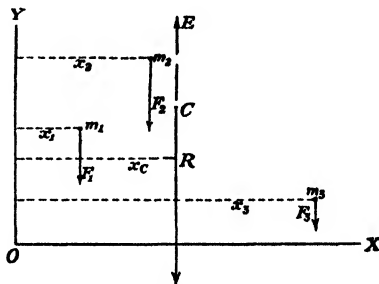


Fig. 33.

If the lever arms x are 1, 2, and 5 ft. respectively, the rule $\Sigma L = 0$ becomes

$$-2 \times 1 - 3 \times 2 - 1 \times 5 + 6x_c = 0, \quad (1)$$

and $x_c = \frac{13}{6}$ ft. This locates the line of action of the equilibrant, and the center of gravity C lies somewhere on that line.

To locate C completely, imagine the body rotated so that the forces act parallel to the X axis as shown in Fig. 34. Let the distance y from the X axis be 2, 3, and 1 ft., respectively. Then applying the second rule, $\Sigma L = 0$, we have

$$2 \times 2 + 3 \times 3 + 1 \times 1 - 6y_c = 0, \text{ and } y_c = \frac{14}{6} \text{ ft.} \quad (2)$$

Thus we have located the center of gravity C , which is $\frac{13}{6}$ ft. from the Y axis, and $\frac{14}{6}$ ft. from the X axis.

If many mass-particles are to be considered, as in a thin sheet of metal or *lamina*, then C is not only the center of gravity but the *center of area* of the surface. Its position is located symbolically

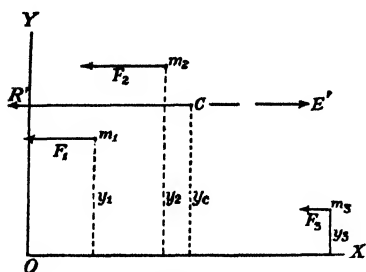


Fig. 34.

as follows: Equation (1) above may be written $\Sigma Fx - x_c \Sigma F = 0$. Therefore

$$x_c = \frac{\Sigma Fx}{\Sigma F}. \quad (3)$$

Similarly equation (2) may be written $\Sigma Fy - y_c \Sigma F = 0$. Therefore

$$y_c = \frac{\Sigma Fy}{\Sigma F}. \quad (4)$$

If the mass-particles extend into three-dimensional space, a third equation is needed, namely

$$z_c = \frac{\Sigma Fz}{\Sigma F}. \quad (5)$$

This locates C with reference to the XY plane.

Since the gravitational forces acting on a body are proportional to the masses of its various particles, the preceding equations may be written in the form $x_c = \Sigma mx / \Sigma m$, and so forth. When expressed in this way x_c , y_c , and z_c locate what is known as the center of mass of a body. It has the same position as the center of gravity, but is defined without reference to any force acting on the mass-particles, and has a meaning similar to the center of population of a state. When the mass-particles are continuous, as in solids, the location of C can be calculated only for relatively simple and homogeneous bodies. Otherwise it can be located only by experiment.

33. Equilibrium of nonparallel forces. When only three forces acting on a body are in equilibrium, their lines of action must either be parallel or meet at a common point. This is because one of them must be the resultant of the other two. But if there are more than three nonparallel forces, they need not meet at a common point. In the case of four, for instance, the resultant of two of them meeting at a point a might be equal and opposite to the resultant of the other two meeting at b . Then if both resultants lie in the line ab , they would be in equilibrium.

Problems involving any number of coplanar nonparallel forces are solved exactly as those concerning parallel forces, but in this case

each force must be resolved into X and Y components. Then there are two systems of parallel forces at right angles to each other, and the general vector equation $\Sigma F = 0$ is split up into two parts $\Sigma F_x = 0$ and $\Sigma F_y = 0$, where the subscripts indicate forces parallel to the X and Y axes respectively. There are then three equations which must be satisfied for equilibrium, namely:

$$\Sigma F_x = 0, \quad (1)$$

$$\Sigma F_y = 0, \quad (2)$$

and $\Sigma L = 0. \quad (3)$

As an illustration, consider a horizontal bar pivoted at O as shown in Fig. 35 (a), with its own weight w acting at its center, and a weight W hung at the end at a distance l from O . A cord supporting it makes an angle θ with the bar. If w alone is taken into account, the lines of action of the forces T (tension in the cord), R (reaction on the pivot), and w must intersect at the common point P as shown in Fig. 35 (b).

Therefore if the force w is produced to intersect the cord, a line from O to this point gives the direction of R . Its magnitude, in this case equal to T , is found from $R \sin \theta = w/2$, since the force triangle is isosceles.

But when W is taken into account, the case is more complicated. It is best solved as follows: The reaction R and the tension T may be resolved into horizontal and vertical components as shown in Fig. 35 (a).

Then from equation (1), $b - d = 0$, and from (2), $a + c - w - W = 0$, where the negative signs indicate forces acting toward the left or downward. Equation (3) taken around O as the center of moments gives $cl - Wl - wl/2 = 0$.

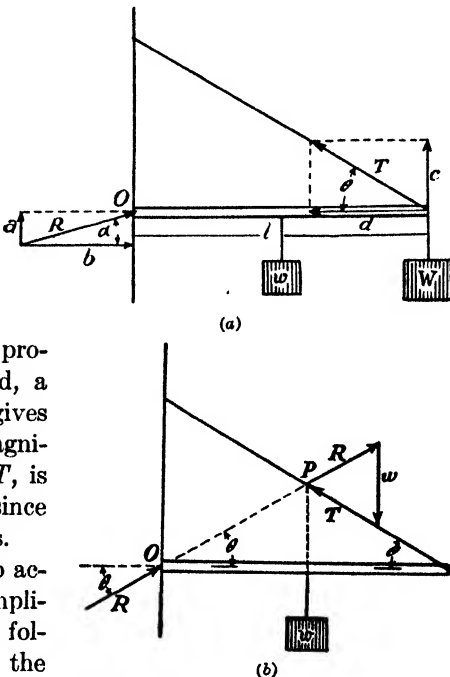


Fig. 35.

Then if w , W , l , and θ are known, the other quantities are found from

$$R = \sqrt{a^2 + b^2}, \quad T = \sqrt{c^2 + d^2}, \quad \text{and} \quad \alpha = \tan^{-1} a/b.$$

The computation of the preceding problem is easy. Suppose the bar weighs 20 lb. and is 12 ft. long. Let a weight of 30 lb. be hung at the end, and let θ be 30° . Then $a = R \sin \alpha$, $b = R \cos \alpha$, $c = T \sin 30^\circ$, and $d = T \cos 30^\circ$. From equation (1), $R \cos \alpha = T \cos 30^\circ$. From (2) $R \sin \alpha + T \sin 30^\circ - 20 - 30 = 0$; therefore, since $\sin 30^\circ = \frac{1}{2}$, $R \sin \alpha = 50 - T \times \frac{1}{2}$. From (3), $T \times \frac{1}{2} \times 12 - 30 \times 12 - 20 \times 6 = 0$. Therefore $T = 480 \div 6 = 80$ lb. Substituting this value in (2) gives $R \sin \alpha = a = 10$ lb., and from (1), $R \cos \alpha = b = 80 \times 0.866 = 69.28$ lb. Therefore $R = \sqrt{a^2 + b^2} = 70$ lb. very nearly. The angle $\alpha = \tan^{-1} a/b = \tan^{-1} \frac{10}{69.28} = 8.2^\circ$. This shows that the weight W shifts the direction of R toward the horizontal. The angle α is equal to θ in Fig. 35 (b), where only w is considered, and it is 0° with only W acting, as was shown in Fig. 24. Therefore with both weights allowed for, the direction of R lies between these limiting values.

SUPPLEMENTARY READING

- J. B. Reynolds, *Elementary Mechanics* (Chap. 3), Prentice-Hall, 1928.
C. S. Whitney, *Bridges; a Study in their Art, Science and Evolution*, Rudge, 1929.

PROBLEMS*

1. A boy weighing 40 lb. sits in a swing and is pulled sideways with a horizontal force of 25 lb. What is the tension in each supporting rope? What is the angle they make with the vertical? *Ans.* 23.6 lb.; 32° .
2. If the ropes in Problem 1 are 10 ft. long, how much force is needed to pull the boy a horizontal distance of 6 ft.? *Ans.* 30 lb.
3. A derrick arranged like the bracket diagram of Fig. 21 lifts a weight of 3000 lb. If the angle the boom makes with the vertical mast is 40° , calculate the tension B in the horizontal supporting rope and the reaction A of the boom, neglecting its own weight. *Ans.* 2517 lb.; 3916 lb.
4. A rope 12 ft. long is fastened to hooks 10 ft. apart at the same level. A weight of 50 lb. is hung from its center. What is the force acting on each hook? *Ans.* 45.2 lb.
5. A 40 ft. boom is supported horizontally at its outer end by a rope from the top of a 50 ft. mast. Calculate the tension on the rope when a weight of 800 lb. is hung from the same end, neglecting the weight of the boom. What is the thrust of the boom? *Ans.* 1025 lb. tension; 640 lb. compression.

* Problems marked with an asterisk are more difficult or more laborious than the average problem in this book.

6. In Problem 5 the boom weighs 400 lb., and if it is of uniform diameter, its weight may be regarded as acting at its center. What is the turning moment of the boom? What is the upward force at the further end of the unloaded boom? What is the total tension on the rope due to the boom's weight and the load at its end? *Ans.* 8000 ft. lb.; 200 lb.; 1281 lb.

7. A crane holds a weight of 100 lb. at the end of a horizontal boom 10 ft. long which is supported by a rope making an angle of 30° with the boom. Calculate the tension in the rope due to the weight, and the longitudinal thrust of the boom. What is the tension if the boom's weight is 50 lb.? *Ans.* 200 lb.; 173 lb.; 250 lb.

8. If the boom of Problem 5 is raised to make an angle of 30° with the horizontal, calculate the turning moment due to its own weight and the 800 lb. load. *Ans.* 34,600 ft. lb.

9. In Problem 8 calculate, by equating turning moments, the tension on the rope which supports the boom. *Ans.* 916 lb.

10. A roof truss supports a weight of 500 lb. as shown in Fig. 23. The pitch of the roof is 37° . Calculate the compression in the beams, and the tension in the tie rod. *Ans.* 415 lb. and 332 lb., nearly.

11. Let five parallel forces as in Fig. 30 have the following values: $F_1 = +2$ lb., $F_2 = -6$ lb., $F_3 = +7$ lb., $F_4 = +4$ lb., and $F_5 = -3$ lb. Their distances from F_1 are 3, 5, 6, and 10 ft. respectively. Calculate the resultant and the distance of its line of action from F_1 . *Ans.* 4 lb.; 2.75 ft. to the right of F_1 .

12. Two men A and B support at its ends a bar 8 ft. long which carries a weight of 50 lb. hung 3 ft. from A . If the bar's weight is neglected, how much does each man carry? Where should the weight be hung so that A may support twice as much as B ? *Ans.* A supports $31\frac{1}{4}$ lb.; B $18\frac{3}{4}$ lb.; $2\frac{3}{4}$ ft. from A .

13. If the bar in Problem 12 is uniform and weighs 12 lb., and if it supports 60 lb. 2 ft. from A , what weight does B carry? *Ans.* 21 lb.

14. Where must a seesaw 12 ft. long be supported if the boys on its two ends weigh 45 lb. and 60 lb. respectively, and if the weight of the board is neglected? If the board is of uniform section and weighs 40 lb., what is the proper point of support? *Ans.* $5\frac{1}{7}$ ft. from the heavier boy; 5.38 ft. from the same.

15. A uniform horizontal bar 10 ft. long and weighing 20 lb. is loaded at each end with 6 and 12 lb. weights, and is pulled upward with a force of 8 lb. at a point 3 ft. from the heavier end. Where must it be supported for equilibrium? *Ans.* 5 ft. $5\frac{1}{2}$ in. from the lighter end.

16. A horizontal boom 8 ft. long and weighing 30 lb. is pivoted at one end, and a weight of 120 lb. is hung at the other end. How much force must be exerted to support the boom at a point 3 ft. from the fulcrum by a rope making an angle of 30° with the mast? *Ans.* 415.7 lb.

17. In Problem 16 calculate the vertical and horizontal components of the tension in the rope by taking moments about the fulcrum. Calculate the net vertical force at the fulcrum. Calculate the resultant reaction of

the mast at the fulcrum and its direction. *Ans.* 360 and 208 lb.; 210 lb.; 295.6 lb. inclined at $44^{\circ}7'$ to vertical.

18. The sides AB and BC of a very thin rectangular slab or "lamina" $ABCD$ are 4 and 6 ft. long respectively. It is loaded at the four corners with weights as follows: 2 lb. at A , 8 lb. at B , 6 lb. at C , and 10 lb. at D . Disregarding the weight of the lamina, find the center of gravity. *Ans.* 2.154 ft. from AD edge; 2.308 ft. from CD edge.

19. A triangular slab ABC has the following dimensions: $AB = 5$ ft., $BC = 4$ ft., and $CA = 3$ ft. It is loaded at A with 4 lb., at B with 8 lb., and at C with 12 lb. Neglecting the slab's weight, find the center of gravity. *Ans.* 6 in. from BC edge; 1 ft. 4 in. from AC edge.

20. An equilateral triangular lamina of uniform thickness is 6 ft. on each side and weighs 12 lb. It is loaded at the corners with weights of 3, 4, and 5 lb. Locate the center of gravity. (NOTE: If unloaded, the centroid lies on the intersection of the median lines.) *Ans.* 1.52 ft. from the 5, 4 edge, and 1.95 ft. from the 3, 4 edge.

21. An equilateral triangular lamina of uniform thickness and 6 ft. on a side is placed above the side of a square lamina in the same plane, having the same length of edge and the same uniform thickness, thus making a figure like the front of a house. Calculate the position of the center of mass. *Ans.* On the median line 4.43 ft. from the base.

22. The bar shown in Fig. 35 weighs 50 lb., is 16 ft. long, and has a weight of 40 lb. hung at the end. The angle between the supporting cord and the bar is 60° . Calculate the tension on the cord, the reaction at O , and the angle it makes with the bar. *Ans.* $T = 75.06$ lb., $R = 45.1$ lb., and $\alpha = 33^{\circ}7'$.

*** 23.** A uniform bar AB is 12 ft. long and weighs 10 lb. It is supported horizontally by cords fastened to its ends. The cord at B makes an angle of 30° with the vertical. A weight of 20 lb. hangs from a point on the bar 4 ft. from end A . Calculate the tension in the cord supporting that end, and the angle it makes with the vertical. *Ans.* 19.53 lb.; $20^{\circ}11'$.

*** 24.** A bar AB is 6 ft. long and weighs 18 lb. It is supported at its ends by cords attached to hooks 10 ft. apart in the ceiling. The cord supporting the end A is 3 ft. long, and the cord supporting B is 4 ft. long. What weight hung from the end A will make the bar hang horizontally? *Ans.* 14 lb.

*** 25.** A tapered bar AB is 10 ft. long and hangs horizontally from two cords fastened to its ends. The cord supporting end A makes an angle of 30° outward from the vertical. The cord supporting B makes an angle of 45° outward from the vertical. Locate the bar's center of gravity. *Ans.* 3.66 ft. from A .

CHAPTER 3

Kinetics

34. Force and motion. Suppose that with no knowledge of physics you were asked to observe the effect of a push or a pull on a body free to move. You would of course notice that in general things remain at rest until you push them, and then that they tend to move in the direction of your push. But you would also notice that once moving they often keep going, for a time at least. You might also find out that it takes a harder push to start a heavy body than to start a light one, and that once started it is harder to stop it. Then if you experimented with, let us say, a lawn roller, you would discover not only that a uniform pull starts it going, but that it steadily gains in speed for a while and that it gains speed faster the harder you pull. Finally, if you were very observing, you might wonder why you were unable to produce motion when you pushed against a stone wall, and you might be clever enough to argue that if the wall were another man he would have to push just as hard as you were pushing in order to keep his position. Then you might say that the wall must be doing the same thing in spite of being devoid of life. In the Dark Ages this amazing doctrine would probably have got you hanged as a sorcerer, but your idea would have been true nevertheless.

The conclusions just arrived at are that bodies at rest need a force to start them, and moving bodies need a force to stop them; that heavy bodies and rapid increases in speed call for stronger forces, and that even inert bodies push back as if they were alive and resented coercion.

35. Newton's laws. All the observations just enumerated were probably made in a vague sort of way thousands of years ago, but it was not until modern times that anyone realized that they could be stated in perfectly general and simple language which applies to any kind of force and to all bodies alike. Such generalizations are usually called "laws." But nature really does not "obey" laws imposed upon her from outside. A law in science is only a man-made

generalization which describes the conditions under which nature has her being, or it expresses a very high probability of what will happen in certain circumstances. It takes a genius to formulate great generalizations, and when Sir Isaac Newton in 1687 first published his three "laws of motion," he was rightly hailed as the greatest of all men of science. These laws, translated from the original Latin, are:

(1) *Every body continues in its state of rest or uniform motion in a straight line, unless it is compelled to change its state by (the action of) impressed forces.*

(2) *Change of motion is proportional to the impressed motive force, and takes place in the direction of the straight line in which that force is impressed.*

(3) *To every action there is always an equal and contrary reaction; or, the mutual actions of two bodies are always equal and oppositely directed.*

36. Discussion of the first and second laws. These two laws are closely related. In fact the first is really a special case of the second. Taken together they express a profound truth derived from experience, that we may measure forces by the *change of motion* which they produce. In his treatise Newton explained that by **motion** he meant what is now called **momentum**, or total quantity of motion. A large mass moving with a certain velocity has a greater quantity of motion than a small mass moving with the same velocity, while with equal masses moving at different velocities, the faster one has the greater "motion." Therefore *momentum depends both upon mass and velocity and is measured by their product, mv .*

The substitution of *momentum* for *motion* in Newton's second law does not go far enough in explaining its meaning. Evidently the change of momentum produced by a given force depends upon how long the force acts. This fact is one of everyday experience. A locomotive takes time to get a train going fast. If it exerts a certain definite pull for say five seconds, the train might attain a velocity of eight miles an hour, but in ten seconds (ignoring friction) it would be going twice as fast. This means that a force F acting for a time t produces an effect which depends upon both F and t , or their product Ft . This product is known as the **impulse** of the force.

We may now restate the second law in the words of Maxwell. *The change of momentum of a body is proportional to the impulse which produces it, and is in the same direction.* Then, if a force F acts upon a mass m , increasing its velocity from v_1 to v_2 , the change in momen-

tum, if m is constant, is $mv_2 - mv_1$, and it follows from the second law that $mv_2 - mv_1 \propto Ft$. This may be written

$$\frac{m(v_2 - v_1)}{t} = kF, \quad (1)$$

where k is a constant of proportionality. The second law thus stated reads: *The time rate of the change of momentum varies as the impressed force.*

37. Definition of the force unit. The second law does not tell us what unit is to be used in measuring force. In fact it leaves us free to choose a unit of any size, because it states only the fact that the time rate of change of momentum *varies* as the impressed force. It does not say that it equals the impressed force. If we wish to measure force in terms of the force-pounds used in the last chapter, and at the same time express m in pounds and v in feet per second, we are quite at liberty to do so. It is necessary to find only the corresponding value of k . As will be explained in the next chapter, this value is 32.174 or approximately 32.2. But if we prefer to adopt the force-kilogram (that is, weight of one kilogram) as our unit, and to measure m in kilograms and v in meters per second, then the corresponding value of k is 9.8 approximately.

A force unit more convenient than either of those just discussed is one which reduces k to unity. This unit is defined as *that force which is just able to cause unit change of momentum in a second*. Then if we assume that the mass remains constant at different velocities, $m(v_2 - v_1)/t = 1$, $F = 1$, and k must then be unity also. The force equation now reads

$$\frac{m(v_2 - v_1)}{t} = F, \quad (1)$$

where F is expressed in terms of this new unit. In the British system of pounds and feet, it is called the **poundal**. In the c.g.s. system, in which m is measured in grams and v in centimeters per second, the force which causes unit change of momentum per second is called the **dyne**, from the Greek word *dynamis* meaning force. This is a very small quantity. It takes 13,825 dynes to equal a poundal, and 4.45×10^8 to equal a force pound; therefore a **megadyne**, or one million dynes, is frequently used to measure large forces.

38. Force and acceleration. Equation (1) in Article 36 may be simplified by setting a for $(v_2 - v_1)/t$, which is the acceleration of the moving body. Then

$$F = \frac{ma}{k}. \quad (1)$$

This is an extremely useful form of Newton's second law, and expresses the fact that when a body's velocity changes under an impressed force, this force is measured by the product of the mass of the body and the resulting acceleration. The force unit defined on this basis is that force which is able to produce unit acceleration of a unit mass. Such a definition makes k equal to unity as before. But this definition of force is not valid unless m is constant while the body speeds up under the action of F . At speeds ordinarily met with, this assumption may be taken as true, but it is now known that the mass of *very* rapidly moving bodies, like the electrons in cathode rays, increases with the velocity, and a constant force applied to such a body does not result in a constant acceleration. The acceleration diminishes as the mass increases. Therefore the definition based on momentum is decidedly preferable, because the time rate of change of momentum does vary directly with the force at all speeds, and is not altered by changes in the mass.

Momentum, being the product of mass and velocity, has the dimensions $[MLT^{-1}]$. Force, which is measured by the time rate of change of momentum, has the dimensions $[MLT^{-1}]/[T] = [MLT^{-2}]$, which are those of mass times acceleration.

39. Impulse. The fact that the impulse Ft of a force is equal to the acquired momentum $m(v_2 - v_1)$, is an aspect of Newton's second law which is very useful in solving certain kinds of problems. Although the word *impulse* is ordinarily used only for short applications of a force like the blow of a hammer, the equation is of course true for any length of time during which a constant force is applied to a mass m . It is especially useful in answering the question, how long must a given force be applied to produce a given change in velocity of a moving mass?

40. Problems involving Newton's second law. As has been explained, this law may be expressed either in terms of change of momentum, by

$$m(v_2 - v_1) = kFt, \quad (1)$$

or in terms of acceleration by

$$ma = kF. \quad (2)$$

The constant k becomes unity when the force is measured in the so-called absolute force units, the dyne and poundal, but it equals 9.8 or 32.2 when the force is measured in kilograms or pounds, and the other units correspond, as specified in Article 37. We are then able to calculate one of the four variables in equation (1) when the three

others are given, or one of the three variables in (2) when the two others are given.

As an illustration, suppose it is required to find how long it will take a force of 80 lb. to stop a mass of 500 lb. moving with a speed of 15 miles an hour. This speed equals 22 ft./sec., which is the total change of velocity $v_2 - v_1$. The constant k is 32.2 since F is given in force pounds; therefore solving (1) for t and substituting the values, we obtain $t = (500 \times 22)/(80 \times 32.2) = 4.27$ sec., very nearly. If the force had been 80 *poundals* instead of pounds, then $k = 1$, and the time would be 137.5 sec.

41. Conservation of momentum. Newton's first law is really a special case of the second, for if the time rate of change of momentum is proportional to the impressed force, then when no force acts, there is no change in momentum.

This constancy of momentum in the absence of an impressed force is a most important property of matter. But one must not infer that the reverse is also true, and that when the momentum of a system does not change there is therefore no impressed force, for there might be several balanced forces acting on it. Therefore if the momentum is constant, either there is no impressed force, or several balanced forces hold the system in equilibrium.

42. Inertia. The remarkable property of matter just explained, is what is known as its *inertia*. Anything having mass has inertia. In fact, the two terms really mean the same thing, and are inherent in what we know as matter, regardless of where it may be found. Thus the inertia, or mass, of a pound of matter on the moon is exactly the same as on the earth, though it would *weigh* very much less there.

This resistance in matter to change of motion, which we call inertia, may be regarded as its most fundamental and significant characteristic. It is inertia which makes it difficult to get a loaded freight car moving, and again it is inertia which makes it hard to stop it. Therefore inertia is just as much a property of bodies at rest as of bodies in motion.

43. The third law—static systems. The word "action" used by Newton in stating his third law may mean either force or the impulse Ft of that force. When no motion results from an action, the time is of no especial significance, so that we then consider only pure balanced forces. There must be at least two forces wherever there is any force at all, which means there can be no force without an opposing force. If you wish to test your strength you must push or

pull against something which resists. The resisting force is exactly equal to the force you exert, and is measured in the same units. Suppose two men wish to break a string by pulling at opposite ends, each exerting a force of one hundred pounds. The tension in the string is one hundred pounds just as it would be if tied to a tree and one man exerted the same force as before. But if both men join forces against the tree, the total tension becomes two hundred pounds, for the tree's reaction is double what it was before.

Action and reaction involve the *mutual action between two bodies*, so that balanced forces acting on a *single body* are not examples of Newton's third law. Thus the men, referred to above, who were pulling on the two ends of a rope illustrate balanced forces but not action and reaction, because they exert equal forces on one body, the rope. But the forces between either man and the rope represent mutual action between two bodies. They are action and reaction, and illustrate the third law.

44. The third law—moving systems. When a locomotive starts a heavy train from rest, one realizes that it is exerting its utmost pull. The noisy exhaust means that steam at full pressure is being admitted to the cylinder throughout the stroke. Then as normal speed is approached the steam is "cut off" earlier in the stroke and the average force it exerts is much less than before. Why should the train's reaction decrease in this remarkable way? The answer calls for an extension of the meaning of the word *reaction*. Until the train is actually moving, the pull exerted by the locomotive is opposed by an equal and opposite pull caused by friction. If there were no friction whatever, the slightest force would cause motion, but since friction exists, a certain definite pull is needed to overcome it. Then the train starts and gains speed. During this process, a reaction due to the train's inertia develops, and according to Newton's second law it takes a force equal to the product of mass and acceleration to overcome it. This is called the **kinetic reaction**. As the train approaches full speed, the acceleration decreases and with it the kinetic reaction, until it is running at a steady rate when the acceleration is zero and the only forces opposing motion are those of rolling friction and the pull of gravity if there is an upgrade.

Newton's third law, extended to include the effect of acceleration, means that the kinetic reaction must be added to other forces which oppose motion. Thus, the force exerted by a locomotive equals the sum of the forces opposing it, whether it is at rest, **gaining speed**, or moving at a constant rate. If f represents the force of friction

when the train is moving on a level track, and m is its mass, the total force needed to give it an acceleration a is found from

$$F = ma + f. \quad (1)$$

The reality of kinetic reaction ma may be shown by the following simple experiment: Let a mass m hang from a spring balance B which registers a force w . This must be equal and opposite to the downward pull of the weight. There are then two balanced forces in equilibrium and no motion takes place if P supports the system. Now let an increased upward pull F cause both balance and weight to move so as to attain a steady velocity. If the balance is carefully observed during this process, it will be found that it records an increased tension while the motion is accelerated. This increase is the kinetic reaction ma . Afterward, with a steady upward motion, the tension comes back to its original value w .

Thus the equation representing all the actions and reactions within the system of balance and weight when accelerated upward, reads

$$F = ma + w. \quad (2)$$

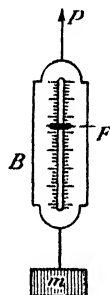


Fig. 36.

As soon as the motion is steady, a vanishes, and we recover the original force w . This shows us that the tension on the ropes supporting an elevator is the same whether it is hanging at rest, or ascending rapidly between floors with a constant velocity. It is, however, much greater during the time it is being accelerated upward, and correspondingly less when it starts to drop, for then ma has a negative sign, as we shall see in the next chapter, where the effect of acceleration on a passenger is fully discussed.

Equations (1) and (2) may be written in the form $F - ma - f$ (or w) = 0, and as each of these terms is an "action" in the broad sense of the word, we may make the general statement

$$\Sigma A = 0, \quad (3)$$

which represents symbolically the fact that the vector sum of all actions to which a body is subjected is always zero. This is true whether it is at rest, gaining or losing speed, or moving at a constant rate, and it is a fundamental principle of mechanics.

As an illustration of this principle, suppose a man is raising a pail of mortar by means of a rope running over a pulley. The pail weighs 80 lb. and the man exerts a pull of 100 lb. What is the upward

acceleration of the pail? Here are three actions: the man's downward pull F , the force of gravity w which opposes it, and the kinetic reaction ma of the pail acting in opposition to its motion. Then $F - w - ma = 0$. Expressing all forces in poundals with k taken as 32 approximately, we have $F = 100 \times 32$ poundals, $w = 80 \times 32$ poundals, and the kinetic reaction is $80a$ poundals. This gives $80a = 3200 - 2560$, and $a = 8$ ft./sec².

45. Equilibrium of momenta. In his statement of the third law, Newton used the words *action* and *reaction* to include not only force, but also impulse and momentum. Thus in equation (2) of Article 44, we may multiply through by the time during which an elevator is being accelerated. Then

$$Ft = mat + wt,$$

or

$$(F - w)t = mat.$$

But the rate of gain of velocity multiplied by the time during which it increases gives the total gain. Therefore $at = v$, and mv represents the gained momentum, or

$$(F - w)t = mv. \quad (1)$$

This indicates that the impulse of the excess upward force, $F - w$, results in an increased momentum. Equation (1) may be generalized for a system of bodies to read

$$t\Sigma F = \Sigma mv, \quad (2)$$

where t is the time during which all the forces act.

If none of the forces indicated by ΣF is applied from outside of the system, it is said to be *isolated*. This would be true of a football game played on a very large raft. As has been explained, ΣF is always equal to zero; therefore Σmv is zero also. This means that the raft and players have neither gained nor lost momentum during the game. Stated in general terms, *it is impossible to alter the momentum of a system from within*. This principle is known as the *equilibrium of momenta* and is a consequence of both the second and third laws of motion. It is really an extension of the conservation of momentum, mentioned in Article 41, to cover cases where forces act *within* a system to produce motion of its parts, but without altering the momentum of the whole.

To illustrate problems involving changes in momentum caused by an impulse, let us suppose that the man pulling up the pail of mortar (Article 44) exerts his hundred pound pull for half a second. Required, the final velocity of the pail. Here two opposing forces, the

pull of 3200 poundals and the opposing force of gravity, 2560 poundals, act on the pail for half a second. The net impulse, $t\Sigma F$, is $\frac{320}{2} = 160$ poundal seconds. This equals the gained momentum mv ; therefore as m is 80 lb., $v = \frac{160}{80} = 2$ ft./sec. This is a rather obvious result, since we found that the acceleration is 8 ft./sec²; therefore the velocity gained in half a second must be 4 ft./sec.

46. Illustrations of the equilibrium of momenta. You cannot make a sailboat go faster by pushing against its sail! This is because your forward push is just balanced by the backward push of your feet on the deck. However, if a man runs from bow to stern of a boat becalmed, he may actually cause it to move forward a short distance. While he is gaining momentum backward, the boat gains an equal momentum forward. These remain constant as long as the man runs at constant speed. Then when he stops, the two momenta become zero once more. The boat is a little ahead of where it was, but the runner is nearly the boat's length behind his original position.†

When an automobile starts forward from rest, the earth gains an equal momentum backward. If it did not the wheels would simply spin around as on ice. If countless cars started to travel along the equator from west to east, they could actually slow down the earth's rotation on its axis, and make the day longer while they continued moving.

Another illustration of equilibrium of momenta is that of a gun firing a shell. Let m be the mass of the shell and M that of the gun. Let v be the shell's velocity and $-V$ the gun's unresisted velocity of recoil. Then, since we may regard the system as isolated, $mv - MV = 0$, both before, during, and after firing. If the gun weighs one thousand times as much as the shell, obviously the velocity of recoil is one thousandth part of that of the projectile.

A shell bursting in mid-air has a certain momentum just before it explodes. Then the fragments fly in all directions, each with a momentum of its own. But their vector sum just after the explosion is the same as that of the shell just before. This means the mathematical center of mass of the flying fragments continues on its original course as if nothing had happened.

Other illustrations of this principle are the rocket and rotary lawn sprinkler. Both depend upon recoil momentum equal and opposite

† In this discussion the effect of friction between water and the boat is ignored. Actually it is possible to get a rowboat moving slowly forward by shifting the rower's weight toward the bow so gradually as not to cause motion of the boat; then a quick movement toward the stern sends the boat ahead.

to the momentum of the ejected fluid. In spite of a common notion, neither device needs air "to push against" and might operate even better in a vacuum.

SUPPLEMENTARY READING

H. A. Erickson, *Elements of Mechanics* (Chap. 7), McGraw-Hill, 1927.

PROBLEMS†

1. A garden roller whose mass is 100 kg is started from rest and reaches a speed of 1.6 m/sec. in 8 sec. What is the acceleration? What force was applied? *Ans.* 20 cm/sec²; 2 million dynes (or 2 megadynes).

2. An automobile whose mass is 3000 lb. speeds up from 14 ft./sec. to 50 ft./sec. in 6 sec. What force in poundals was applied? In pounds? *Ans.* 18,000 poundals; 559 lb.

3. A force of 32×10^6 dynes is applied to a mass of 40 kg. What speed does it attain in 2 minutes? *Ans.* 960 m/sec.

4. A force of 5 kg is applied to a mass of 12 kg moving against it with a speed of 8 m/sec. How long will it take to stop it? *Ans.* 1.96 sec.

5. A force of 80 lb. stops a handcar going at 12 miles per hour in 15 sec. What is its mass? *Ans.* 2195.4 lb.

6. How great a force in pounds is needed to produce an acceleration of 8 ft./sec². in a mass of 20 lb? How many poundals? *Ans.* 4.97 lb.; 160 poundals.

7. How much force is required to produce an acceleration of 200 cm/sec². in a mass of 4 kg against a force of 980 dynes on each gram? *Ans.* 4.72 megadynes.

8. What acceleration is produced on a mass of 100 lb. when pulled upward with a force of 600 lb.? *Ans.* 161 ft./sec².

9. A rifle weighing 6 kg fires a bullet weighing 30 g. Its muzzle velocity is 700 m/sec. What would be the velocity of the gun's recoil if it were perfectly free? *Ans.* 350 cm/sec.

10. In Problem 9, if the acceleration were uniform through the barrel, and the time taken in reaching the muzzle is 0.003 sec., what is the force of the recoil? *Ans.* 700 megadynes.

11. Calculate the horizontal recoil velocity of a 400 lb. gun mounted on wheels, which shoots a projectile weighing 20 lb. with a velocity of 2000 ft./sec. at an angle of 30° with the horizon. *Ans.* 86.6 ft./sec.

12. A man weighing 200 lb. runs forward along the deck of a 3000 lb. boat lying becalmed in still water. The runner's speed is 16 ft./sec. measured along the deck. What is the backward velocity of the boat through the water? *Ans.* 1 ft./sec.

† In these and all similar problems in this book, the acceleration is assumed to be constant. Friction is for the present neglected.

13. If the boat in Problem 12 is 28 ft. long, how far does it move backward, assuming that the man reaches full speed instantaneously at the stern and stops instantaneously at the bow? *Ans.* 1 ft. 9 in.

14. Show that the above result follows from the law of the conservation of momentum of a self-contained system. The center of mass of boat and runner remains fixed in space regardless of how he reaches the bow, and the distance the boat moves is always the same.

CHAPTER 4

Gravitation and Falling Bodies

47. Gravitation. In addition to his laws of motion, Newton, in 1672, formulated the law of "universal gravitation." This was derived from a study of the moon's orbit and showed for the first time what was the underlying principle of the solar system. Newton proved that the known behavior of the planets and moon in their orbits could be explained by assuming a force which urges them together, and he made the startling statement that the moon is held in her orbit by the same kind of force which makes the apple fall from the tree. Both are illustrations of the universal attraction which Newton called gravitation, and which exists between any two masses wherever situated. Both are *mutual* attractions, for the moon and the apple pull the earth just as strongly as the earth pulls them.

According to Newton's law, the gravitational attraction between two mass particles is directly proportional to the product of their masses and inversely proportional to the square of the distance between them. In algebraic language, this reads

$$F \propto \frac{m_1 m_2}{r^2}, \text{ or } F = \frac{G m_1 m_2}{r^2},$$

where G is the constant of proportionality to be determined.

If we have many particles forming two homogeneous spheres, r is the distance between their centers, because it can be proved that either a solid sphere or hollow spherical shell acts as though all its mass were concentrated at its center.

In stating his formula, Newton did not know the value of G because his observations were limited to the gravitational pull of the earth and other bodies of the solar system, none of whose masses was known. He knew the force F with which the earth attracts a mass m_2 , and in addition the distance r to the earth's center, but G remained unknown to him because the mass of the earth (m_1) had not yet been determined.

In 1797 and 1798, Henry Cavendish, an English chemist and mathematician, succeeded in measuring the attraction between two pairs of metal spheres in his laboratory. As F , r , and the masses

of the spheres were known, G could be calculated. Then it was comparatively easy to determine the masses of the earth, sun, moon, and planets.

Recent and much more accurate observations by Dr. Paul R. Heyl of the United States Bureau of Standards give $6.664 \times 10^{-8} \text{ cm}^3/\text{g-sec}^2$ as the value of G . Thus gravity is seen to be an extremely small force except when one of the attracting bodies is large, like the earth.

48. Acceleration of falling bodies. As the force which makes the apple fall varies with the apple's mass, it might seem at first sight as if large apples would fall faster than small ones. In fact, this was formerly supposed to be the case. The ancients affirmed that heavy bodies fell faster than light ones. They considered this mere common sense and no one took the trouble to see whether it were true or not. Then Galileo in 1590 performed his famous experiments of dropping objects of different masses from the Leaning Tower of Pisa, and showed that all reached the ground at practically the same time. This demonstration that the ancient philosophers were wrong was considered impious at that time, and contrary to reason. But Newton's second law clears up the difficulty. The force on the heavier mass *is* greater, but it takes more force to accelerate it, so the result is unaltered by the increased mass. If one body has twice the mass of another, the force pulling it downward is twice as great, but it takes twice as much force to produce the same acceleration. Therefore, if dropped together from a tower, they would reach the ground at exactly the same time provided air resistance could be eliminated. Actually, objects of very low density like cork or feathers fall much more slowly in air than dense bodies like lead. This is because air resistance counts more in relation to weight with a falling body of low density like cork, than with a dense body like lead, although each may have the same shape and volume. But in a vacuum a feather falls as fast as a bullet, and then both are called *freely falling bodies*.

The value of the acceleration of freely falling bodies may be calculated from Newton's gravitational equation. Let m_1 be the mass of the earth, and m_2 that of the falling body. The force acting on m_2 to produce acceleration is m_2a ; therefore

$$\frac{Gm_1m_2}{r^2} = m_2a, \quad \text{or} \quad a = \frac{Gm_1}{r^2}.$$

Here r is the earth's radius, so with G and m_1 known, a can be determined. But the earth is not a perfect sphere, and its density is not

uniform, so that the result obtained in this way is not very accurate. If we desire precision, the acceleration due to gravity, designated by g , must be obtained by direct experiment. Its value is found to vary from point to point, and exceedingly delicate measurements are continually being made all over the world to determine this important physical quantity. Its value at the Bureau of Standards, Washington, D. C., is 980.097 cm/sec². and in general it varies between the limits of about 978 cm/sec². at the equator to 983 cm/sec². at the poles. In problems where high precision is not needed, g is ordinarily taken as 980 cm/sec²., or 32.2 ft./sec².

49. The weight of bodies. Weight and mass are not the same thing. Mass is a measure of a body's inertia, and would not be altered if the force of gravitation were to change or even vanish. Weight, on the other hand, is a measure of gravitational attraction. It is generally used to designate the attraction between the earth and a given mass, though the same mass would have weight of a different value on the moon or on any other heavenly body.

Since weight, like other forces, may be measured in accordance with the equation $F = ma$, we may obtain the terrestrial weight of a mass m by substituting for a the particular acceleration g which is due to the earth's gravitational field. Then

$$w = mg, \quad (1)$$

where the force of gravitation w is measured in absolute units and k is unity. Thus the weight, or gravitational attraction, of a pound mass is 32.2 poundals, wherever $g = 32.2$ ft./sec²., and the weight of a gram mass is 980 dynes wherever $g = 980$ cm/sec².

50. The gravitational units of force. When a force of one pound as defined above acts upon a mass of one pound, the acceleration is 32.2 ft./sec²., very nearly. But a force of a poundal by definition accelerates a pound mass at the rate of only one foot per second per second. Therefore, since the force varies as the acceleration, a force-pound is 32.2 times as large as a poundal. We have seen that in the equation stating Newton's second law, $F = ma/k$, the constant k becomes unity when F is expressed in poundals. But if F is to be given in terms of the larger unit, its numerical value must be 32.2 times smaller, just as a distance measured in miles is numerically smaller than the same distance measured in feet. This means that k is no longer unity but has the value 32.2. Thus F (in pounds) = $ma/32.2$, and in the c.g.s. system, F (in grams) = $ma/980$.

It is incorrect to write $F = ma/g$, because F and mg are both

forces so that their ratio, $ma/F = k$, is a pure number without dimensions, and cannot equal g , which is an acceleration by definition.

We have now shown that when absolute force units are used (poundals or dynes), the constant k is unity, and that when gravitational units (force-pounds and grams) are used, it is 32.2 in the English system, and 980 in the c.g.s. system. It should also be noted that when the equation $F = ma/k$ is used, m and a must be measured in corresponding units. When F is in dynes, m must be stated in grams and a in cm/sec². When F is in force-pounds, m must be stated in pounds and a in ft./sec². When F is in force-kilograms, m must be stated in kilograms and a in m/sec²; in the latter case k is equal to 9.8.

51. The slug. This is a unit of mass which is now much used in the United States, although it is called the "British Engineering Unit." It was adopted in order to give k a unit value and at the same time permit the use of the gravitational unit, the force-pound. The **slug** is defined as *that mass which when acted on by a force-pound gains velocity at the rate of one foot per second per second*. Thus in the force equation $F = ma/k$, if we set m equal to one slug, a equal to one ft./sec², and F equal to one force-pound, k must be unity also. But when a force of one pound acts upon a pound mass, the acceleration is 32.174 ft./sec²; therefore, a slug must equal 32.174 pounds if it is to be accelerated at the rate of only one ft./sec². by a force-pound.

If m is expressed in slugs, the force equation becomes $F = ma$, and calculations are as simple as when absolute force units are used. However, reducing pounds to slugs involves dividing m by 32.174 (or 32.2 approximately). This is equivalent to dividing ma by the same quantity when we use gravitational force units. So there is no saving in the numerical computation.†

52. Calculation of accelerated motion. If an automobile starting from rest gains speed at the rate of say 11 ft./sec., it will be going at the rate of 44 ft./sec., or 30 miles per hour, at the end of 4 sec. In general, the velocity gained in the time t is given by $v_t = at$. But if an auto going at 44 ft./sec. is accelerated as above for the same time,

† There is still another method for dealing with the gravitational measure of force. It consists in regarding *force*, distance, and time as the basic units, instead of *mass*, distance, and time. This is perfectly logical when using the British units, where "pound weight" is the legal standard. In such a system, mass becomes a derived unit given by $m = w/g$. Dividing weight in pounds by 32.2 ft./sec². gives mass measured by a new unit numerically equal to the slug, whether we call it that or not.

the final speed would be 88 ft./sec., or 60 miles per hour. This may be expressed by

$$v_t = v_o + at, \quad (1)$$

where v_o means the original velocity, and a may be either positive or negative. A negative acceleration, such as occurs when a car slows down, tends to *decrease* the original velocity and may be called a "deceleration."

Let us next suppose that we know the values of the initial and final speeds of the car and wish to know how far it has moved while speeding up or slowing down from one speed to the other. If we assume that the acceleration is uniform (not necessarily true), then the average velocity, v_a , during the time the speed rose from 44 to 88 ft./sec. is the average of these values, or 66 ft./sec., and in the 4 sec. assumed above, it must have traveled $4 \times 66 = 264$ ft. This conclusion may be expressed in general terms by

$$s = v_a t = \frac{(v_o + v_t)}{2} t. \quad (2)$$

A third important problem requires the distance when the initial velocity, acceleration, and time are given. In the case of the automobile, if the initial velocity is 44 ft./sec., the acceleration 11 ft./sec.², and the time 4 sec., how far does it go? Here we know that without any gain in speed it would go $44 \times 4 = 176$ ft. But due to the acceleration, the speed gains $11 \times 4 = 44$ ft./sec. during 4 sec., or an average gain of 22 ft./sec. The distance covered at this average rate is $22 \times 4 = 88$ ft. So the total distance covered is $176 + 88 = 264$ ft. Summing up these conclusions in symbols, we obtain

$$s = v_o t + \left(\frac{at}{2}\right)t,$$

$$\text{or} \quad s = v_o t + \frac{at^2}{2}. \quad (3)$$

Equation (3) is easily derived algebraically by substituting the value of v_t from (1) in (2), giving

$$s = \frac{1}{2}(v_o + v_o + at)t = v_o t + \frac{1}{2}at^2,$$

as before.

A fourth equation gives the final velocity in terms of its initial value and the distance traveled. It is obtained most easily by eliminating t between (1) and (2). Thus from (1)

$$t = \frac{v_t - v_o}{a}.$$

Substituting in (2) we obtain

$$s = \frac{(v_t + v_o)}{2} \times \frac{(v_t - v_o)}{a} = \frac{v_t^2 - v_o^2}{2a}.$$

$$\therefore v_t^2 = v_o^2 + 2as. \quad (4)$$

53. Motion of falling bodies. The equations derived above apply *exactly* to freely falling bodies, for then a is definitely constant. Therefore, in discussing them we shall substitute g for a in the four equations of Article 52.

An examination of equation (3) reveals some important facts regarding freely falling bodies. If, for example, a body is dropped from rest, v_o is zero, and if we consider downward motion as positive in this case, $s = \frac{1}{2}gt^2$. The total distance traveled, therefore, increases as the square of the time, and if g is taken as 32 ft./sec²., the distance fallen in 1 sec. is $16 \times 1 = 16$ ft., in 2 sec. it is $16 \times 4 = 64$ ft., in 3 sec. it is $16 \times 9 = 144$ ft., and so forth. But the distances fallen through during the successive seconds are (1st sec.) 16 ft.; (2nd sec.) $64 - 16 = 48$ ft.; (3rd sec.) $144 - 64 = 80$ ft., and so forth.

If the initial velocity is not zero, then the fact that g (like a) may be either positive or negative, is most significant. When a body is thrown straight downward with a velocity v_o , it would cover the distance v_ot in t seconds, regardless of gravity. But the acceleration increases the velocity precisely as it would have done if the body had started falling from rest, and the second term, $\frac{1}{2}gt^2$, gives the corresponding increment in s .

If you throw a stone directly upward, the effect of gravity is progressively to diminish its velocity with the negative acceleration g , so that instead of rising at a uniform rate, the body begins to fall, as it were, the moment it leaves the hand. This falling effect ultimately neutralizes the upward velocity, and when this occurs $v_t = 0$. Then equation (1) becomes $v_o = gt$, and may be written $t = v_o/g$. This means that if we destroy v_o at the rate g , it takes v_o/g seconds to destroy it completely. But if the stone is dropped from rest, the velocity gained in the time t is obtained from $v_t = gt$, which is the same as the velocity lost in the same time when it is thrown upward. This means that it would return to the hand with the same speed with which it left it after twice the time required to reach its highest point, a total of $2v_o/g$ seconds.

54. Illustrative examples. Suppose a stone is thrown upward with a velocity of 256 ft./sec. from a point 2000 ft. above a plain. Re-

quired, the height to which it will rise and its velocity and position at the end of 20 seconds. The time it takes to come to rest is $\frac{256}{32} = 8$ sec. It will then be $s = \frac{1}{2}gt^2 = 1024$ ft. above its point of departure. This point is again reached in $2 \times 8 = 16$ sec. from the start, and the stone then has a velocity of 256 ft./sec. downward. Four seconds later (at the end of 20 seconds) it will have fallen a distance $s = v_0t + \frac{1}{2}gt^2 = 256 \times 4 + 16 \times 4^2 = 1280$ ft. below the starting point, which is 720 ft. above the plain. The velocity will then be $v_t = v_0 + gt = 256 + 32 \times 4 = 384$ ft./sec.

It might also be required to find the point at which the stone has reached some assigned velocity. Let this be 64 ft./sec. Then substituting in (4) we have

$$64^2 = 256^2 - 64s, \text{ and } s = 960 \text{ ft.}$$

This is 64 ft. below its highest point, and is also the distance it would fall in 2 sec., so that the stone while ascending reaches this level 6 sec. after it leaves the hand, and 4 sec. later, on its descent. If it is required to find the time it takes to reach a given height, we may use (3) which is a quadratic in t , the double solution referring to the time of passing the given level both on the way up and on the way down. But it is somewhat easier to calculate the time and height of the ascent first, and either add or subtract from it the time it would take to fall the necessary distance. Thus if s is given as 960 ft., we find that it takes 2 sec. to fall $1024 - 960 = 64$ ft., so that the two answers are 6 and 10 sec. as seen above.

55. Projectiles. When a rifle is fired horizontally, the bullet begins to fall the moment it leaves the muzzle, exactly as if it had been dropped from the same point. If the marksman is standing on a level plane, two bullets, one fired and the other dropped at the same instant from the muzzle, would strike the ground simultaneously. The reason we can aim "point blank" at a nearby target is that in the small fraction of a second required for the bullet to reach the mark, the drop is too small to be observed. These two motions, horizontal and vertical, have no influence upon each other, as we saw was the case with vectors in general when perpendicular to each other.

As a simple example in the theory of projectiles, suppose a rifle bullet fired horizontally from a height of 144 ft. above a plane as indicated in Fig. 37. Applying $s = \frac{1}{2}gt^2$, we find that it takes 3 sec. to reach the plane. If the initial velocity were 2000 ft./sec., it would travel horizontally 6000 ft. during that time. This is the range l ,

and its trajectory would be the curve OP . The equation of this curve is obtained by eliminating t from $x = v_0 t$ and $y = -\frac{1}{2}gt^2$, giving $y = -gx^2/2v_0^2$. But $g/2v_0^2$ is a constant, so the equation may be written $y = -cx^2$, where $c = g/2v_0^2$. This is the equation of a parabola passing through the origin as shown in the diagram. The ordinates y_1, y_2, y_3 , and so forth, are proportional to the squares of the corresponding abscissae, because x is proportional to the time, while y is proportional to the time squared. In general, all projectiles would describe true parabolic curves if the effect of air resistance could be eliminated. As it is, their trajectories more or less approximate parabolas.

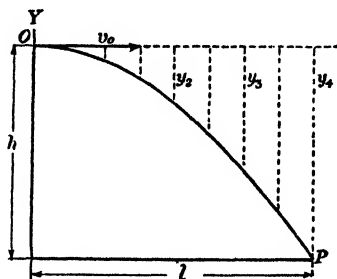


Fig. 37.

When a projectile is thrown at some angle θ with the horizontal, as indicated in Fig. 38, it is necessary to resolve the velocity v_0 into horizontal and vertical components, and then study the motion of each independently. The vertical velocity v_y equals $v_0 \sin \theta$. This takes the place of v_0 in the equations describing accelerated motion, so we may calculate the time it takes the body to reach its highest

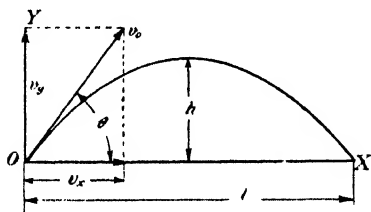


Fig. 38.

point as if it had been thrown vertically upward. The total time T the projectile is in the air is twice the time of rising, which is calculated from $t = v_0 \sin \theta / g$ instead of $t = v_0 / g$. When T is multiplied by the horizontal component of the velocity, $v_x = v_0 \cos \theta$, we get the range l . The

maximum height of the trajectory h is found from the general equation $s = \frac{1}{2}gt^2$, where the height is s , and t the time of ascent.

As a numerical illustration of the preceding problem, let θ be 30° and v_0 be 960 ft./sec. Then the vertical component $v_y = v_0 \sin \theta = \frac{960}{2} = 480$ ft./sec. The total time is given by $T = 2v_y/g = 2 \times \frac{480}{32} = 30$ sec. The horizontal component $v_x = v_0 \cos \theta = 960 \times 0.866 = 831$ ft./sec., and the range $l = v_x T$ is $831 \times 30 = 24,930$ ft. The highest point h is $gt^2/2 = 32 \times 15^2/2 = 3600$ ft.

56. Motion on an inclined plane. If a body slides down a perfectly smooth inclined plane it behaves like a freely falling body, but

with its acceleration diluted, as it were, by the slope. The force of gravity mg may be resolved into two components parallel and normal to the slope. These are $mg \sin \theta$ and $mg \cos \theta$ as shown in Fig. 39,

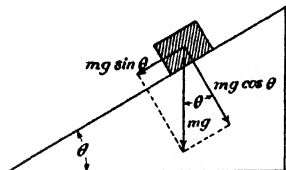


Fig. 39.

where the angles marked θ are equal, as their sides are mutually perpendicular. The normal force has no effect on the motion, but the force $mg \sin \theta$ causes an acceleration $g \sin \theta$, and this quantity is used in place of a in the equations of accelerated motion. Thus we may compute the time of descent, distance covered,

velocities, and so forth, if the initial velocity up or down the plane and the angle θ are known.

57. The elevator problem. In the preceding paragraphs we were concerned wholly with the force of gravity which varies with the body's mass, and with accelerations which were independent of the mass of the moving body. But other forces that may act upon a body have nothing to do with its mass. Then any acceleration is possible, and in any direction in which the body is free to move. Thus an elevator is accelerated upward against gravity when it starts from rest, because a second force greater than that of gravity acts upon it. Applying $\Sigma A = 0$, we obtain for this particular case, $F - mg - ma = 0$, where mg is the weight and ma is the reaction caused by the upward acceleration a . This equation may be written $F = m(g + a)$ which shows that the total downward pull of the elevator is as if the acceleration due to gravity had been increased by the amount a . All objects within it are similarly affected and have their weight apparently increased by ma . If a man "weighing" 160 lb. is standing on a spring balance in the elevator as it starts upward with an acceleration of 4 ft./sec², his apparent weight would become $160(32 + 4) = 5760$ poundals, or $\frac{5760}{32} = 180$ force-pounds, his mass of course remaining constant. But as soon as a steady upward velocity is attained, $a = 0$, and $F = mg$ as before, so that no additional weight would be recorded by the balance.

If the elevator starts downward, a is negative and the man's apparent weight would be 160×4 poundals less than usual, so that the balance would record only 140 pounds. If the downward acceleration were equal to g , then $F = m(g - g) = 0$, and objects within it would apparently weigh nothing, because both they and the elevator would be falling freely through space. If it could be accelerated downward at a still greater rate, the passengers would "weigh"

less than nothing with respect to the floor of the elevator and would "fall" on the ceiling.

58. Forces acting on freely sliding bodies. The force required to pull a body up an inclined plane may be conveniently measured by a device represented in Fig. 40. A small mass m hangs at the end of a string which passes over a light, nearly frictionless pulley P , and is attached to the large mass M . This latter mass rests on a smooth surface inclined at an angle θ with the horizontal. To determine the result of this arrangement, let us apply the action principle, $\Sigma A = 0$, to the system of two masses. First, the weight m gives us a force, mg in absolute units, which tends to move M up the incline. Similarly, the force Mg acting on M has a component $Mg \sin \theta$ urging it down the plane. Finally, unless these two forces are balanced, there is a kinetic reaction caused by the acceleration of both masses. This equals $(M + m)a$, and may be regarded as added to $Mg \sin \theta$ in opposing motion up the plane. Then the action equation becomes

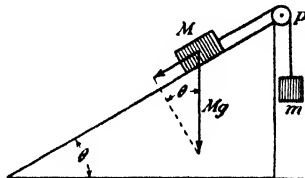


Fig. 40.

$$mg - Mg \sin \theta - (M + m)a = 0,$$

from which a may be found, if the masses and angle are known. If there is no motion, a is zero, and $mg = Mg \sin \theta$, or $\sin \theta = m/M$, by which the critical angle for two given weights may be calculated. If $Mg \sin \theta$ is greater than mg , the mass M slides down the plane and the kinetic reaction opposing this motion must have the same sign as mg , which also opposes it.

If the angle θ in the preceding problem is 90° , the two masses must be equal in order to balance each other. If, however, one has a small excess mass added to it, a small acceleration results. This arrangement is shown in Fig. 41, where one of the two equal masses M has a rider m resting on it. The action equation then becomes

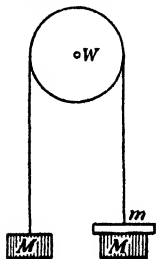


Fig. 41.

$$(M + m)g - Mg - (2M + m)a = 0,$$

whence

$$mg = (2M + m)a,$$

or

$$a = \left(\frac{m}{2M + m} \right) g.$$

As m is supposed small compared to $2M$, a is much smaller than g , and the loaded mass gains speed downward so slowly that the accel-

eration may be measured with some precision. With certain accessories not shown, this device, known as Atwood's machine, may be used to determine g approximately.

If the angle θ is 0° , the surface is horizontal, as shown in Fig. 42, and $mg = (M + m)a$. In this case, if there is no friction, the system

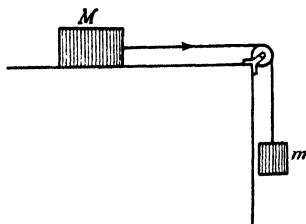


Fig. 42.

is always unbalanced, and motion results. The amount of acceleration depends only upon the applied force mg , and the total mass accelerated $M + m$.

If the force applied to M does not act along the surface, but at an angle θ as shown in Fig. 43, we must resolve F into its components normal and parallel to the plane. The parallel component

$F \cos \theta$ is the effective one, and the action equation becomes $F \cos \theta - Ma = 0$. From this we may obtain the acceleration produced by a known force acting at a known angle on a given mass.

In the preceding problems the speed acquired in a given time might be wanted, or the time it would take to attain a given speed. Then we must use the impulse of the applied force which equals the momentum gained by the system. Thus in the case shown in Fig. 42, the equation multiplied by t becomes

$$mgt = (M + m)at,$$

or

$$mgt = (M + m)v,$$

while for the case of Fig. 43, we use $Ft \cos \theta = Mv$. Thus either t or v may be found if the other is given as well as the masses, or force and angle.

This type of problem is of great practical importance when applied to a locomotive getting a train up to speed, either on a level track or upgrade. If the "draw-bar pull" of the locomotive, the mass of the train, and the force of friction are known, we may calculate the time required to attain full speed on a given grade. But we cannot discuss such examples further until we have found out how to handle friction, which is far from negligible in actual practice. This will be considered in the next chapter.

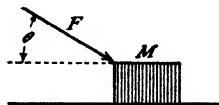


Fig. 43.

SUPPLEMENTARY READING

H. A. Erickson, *Elements of Mechanics* (Chap. 3), McGraw-Hill, 1927.

PROBLEMS†

1. How far does a stone fall in 5 sec. from rest? What is the acquired velocity? *Ans.* 400 ft.; 160 ft./sec.

2. How long does it take a stone to fall 100 ft.? What is the final velocity? *Ans.* 2.5 sec.; 80 ft./sec.

3. A body moving with a constant acceleration has a speed of 12 ft./sec. when passing a given point. Forty-two feet farther on its speed is 30 ft./sec. Calculate the acceleration and the time required to reach the second point. *Ans.* 9 ft./sec²; 2 sec.

4. If the acceleration in Problem 3 were 6 ft./sec²., what must be the distance between the two points where the speeds are as given? *Ans.* 63 ft.

5. A stone is thrown vertically upward with a velocity of 120 ft./sec. How high will it rise? How long will it be in the air? *Ans.* 225 ft.; 7.5 sec.

6. The splash of a stone dropped into a deep well is heard 6 sec. later. Estimate the time of falling with allowance for the speed of sound taken roughly at 1000 ft./sec., and calculate the depth of the well. *Ans.* 5.51 sec.; 490 ft.

7. A stone is thrown vertically upward with a velocity of 78 ft./sec. relative to and from a balloon 400 ft. above the ground. The balloon is rising at the rate of 50 ft./sec. When will the stone strike the ground, and with what velocity? *Ans.* 10.4 sec. later; 204.8 ft./sec.

*8. In Problem 7, when does the stone pass the balloon on the way down? *Ans.* 4.88 sec. later.

9. A stone is thrown horizontally with a velocity of 60 ft./sec. from a cliff 100 ft. high above a plane. How long is it in the air? How far from the base does it strike? What is the final resultant velocity? *Ans.* 2.5 sec.; 150 ft.; 100 ft./sec.

10. A projectile is thrown from a gun at an angle of 40° with a plane, and with a muzzle velocity of 2000 ft./sec. Calculate the time it is in the air, the height to which it rises, and its range. *Ans.* 80.4 sec.; 4.9 miles; 23.32 miles.

11. A shell fired from a rifle is 20 sec. in the air, and its range over a plane is 12 km. What was the elevation of the gun? *Ans.* 9°3.

12. A body slides down a frictionless plane at an angle of 30° with the horizontal. What is its velocity after sliding 6 m from rest? How long does it take? *Ans.* 766 cm/sec.; 1.564 sec.

13. A body sliding down an inclined plane acquires a speed of 4 m per second in 3 sec. What is the slope? *Ans.* 7°8.

14. A body is projected up an incline of 20° with a speed of 20 m/sec. measured along the slope. When will it return to the starting point? How far along the plane does it move? *Ans.* After 11.94 sec.; 59.7 m.

† In this set of problems, use $g = 32$ ft./sec²., when distances are given in feet. Air friction is ignored.

15. How long will it take a force of two tons (4000 lb.) to accelerate upward an elevator weighing 1.8 tons from rest to a velocity of 5 ft./sec.? *Ans.* 1.4 sec.

16. What is the tension in the rope if the elevator of Problem 15 moves downward with the same acceleration? *Ans.* 3201 lb.

17. What downward acceleration must an elevator have if a man weighing 75 kg is to "lose" 15 kg in weight? If the upward acceleration is 150 cm/sec., how much weight does he "gain"? *Ans.* 196 cm/sec²; 11.48 kg.

18. In Atwood's machine the weights are 500 g each. The rider weighs 50 g. Calculate the resulting acceleration and the distance moved in 3 sec. *Ans.* 46 $\frac{2}{3}$ cm/sec., 210 cm.

19. In Fig. 40, the mass m is 6 lb. The mass M is 10 lb. and the slope is 30°. Calculate the acceleration, and the velocity when m has descended 16 ft. from rest. *Ans.* 2 ft./sec²; 8 ft./sec.

20. What is the slope of the above plane, if it takes 10 sec. for m to descend 16 ft.? *Ans.* 35°7.

21. In Fig. 42, m is 200 g. It descends 3 m from rest in 2 sec. What is the mass M ? *Ans.* 1106.7 g.

22. A force of 24 lb. is applied steadily at an angle of 45° to a 200 lb. mass as illustrated in Fig. 43. Calculate the velocity after it has moved 12 ft. from rest. *Ans.* 8.06 ft./sec.

23. What is the backward push in pounds of a sprinter weighing 160 lb. who reaches his full speed of 10 yd./sec. in 8 ft.? *Ans.* 281.25 lb.

24. A force of one million dynes acts on a mass of 50 g for 0.2 sec. What is the acquired velocity? What force in kg is needed to stop it in a distance of 5 cm? *Ans.* 40 m/sec.; 81.6 kg.

25. A bullet weighing 25 g is shot from below into a block of wood suspended by a string. The block weighs 2 kg and rises 38 cm as a result of the impact. What is the velocity of the bullet? *Ans.* 221 m/sec.

* 26. At what angle and what initial speed must a shell be fired to hit an observation balloon horizontally when it is 10,000 ft. above a point 34,642 ft. from the gun? *Ans.* 30°; 1600 ft./sec.

CHAPTER 5

Work, Energy, Power, and Friction

59. Work. Why is it harder work to climb a mountain than to walk the same distance on the level? The reason is that each step has a vertical component, so that we are raising our own weight directly against the opposing force of gravity. In general *any motion against an opposing force calls for work*, and the work done is the product of the force and the component of the distance in line with the force, or the product of the distance and the component of the force in line with the distance moved. In symbolic language this is $Fs \cos \theta$, where θ is the angle between the two vectors F and s , the force and the distance. If the two are in line, $\cos \theta = 1$, and the work is the product of the force and the distance through which it acts.

It is very important to recognize the fact that work is done only when the force results in motion. Holding a heavy weight at arm's length is very fatiguing,[†] and requires great muscular strength, but the muscles are doing no work against gravity, as is evident when we realize that a bronze statue could do this kind of "work" better than we could.

60. Unit and dimensions of work. The unit of work is that accomplished when a unit force acts through unit distance. In the c.g.s. system, where the force is a dyne, and the distance one centimeter, this unit of force-distance is called the **erg**, from the Greek word "ergon," meaning work. But as the erg is extremely small, another unit 10^7 times as large is commonly used. This is named the **joule**, after James P. Joule, whose investigations have greatly advanced our knowledge of thermal energy.

In the English system the corresponding unit is a *foot-poundal*, but this absolute unit is little used, and instead the gravitational **foot-pound** is generally employed. This is the work done when one pound of force acts through one foot. The *gram-centimeter* and *kilogram-meter* are corresponding units in the metric system.

[†] Our fatigue is caused by the myriads of twitching contractions of individual muscle fibers whose overall effect is the lifting force that is exerted. Thus, though no net work against gravity is accomplished, there is a great deal of motional activity in the muscle cells which consumes energy and results in fatigue.

The dimensions of work are obtained by taking the product of those of force and distance, so that

$$[W] = \left[\frac{ML}{T^2} \right] \times [L] = [ML^2T^{-2}].$$

61. Energy. We call a man energetic if he is able to accomplish a great deal of work. He possesses a quality known as *energy*. It means literally "in-work," and may be the source of either mental or physical activity. But in mechanics, *work* refers only to forces acting through distances, and *energy* means only *the capacity for, or ability to do mechanical work*. Energy, then, must be measured in terms of work. Therefore it has the same unit and same dimensions.

These two ideas, work and energy, are related in much the same way as action and reaction, or buying and selling, where the same quantity, pounds or dollars, measures both aspects of the same transaction.

62. Power. It takes a more *powerful* man to pull up a heavy anchor by a windlass in five minutes, than one who can just do it in ten. Everyone recognizes the obvious fact that a man can do as much work as a horse if he is given time enough. By means of a block and tackle he can move the heaviest vehicle. But to do so as quickly as a horse, demands the horse's power. Therefore **power** involves the notion of time as well as work, and is defined as *the time rate at which work is done*. In symbols, power is expressed by $P = W/t$, and as the dimensions of W are $[ML^2T^{-2}]$, the dimensions of P are

$$\left[\frac{ML^2T^{-2}}{T} \right], \text{ or } [P] = [ML^2T^{-3}].$$

63. Units of power. When work is done at the rate of an erg per second, the power is the absolute unit in the c.g.s. system. This unit has no especial name, because it is too small for common use. But 10^7 ergs per second, or a joule per second, is called a **watt**, after James Watt, the celebrated engineer. One thousand watts is known as a **kilowatt**, a unit much used by electrical engineers.

In the English system, the absolute unit would be a foot-poundal per second. But the unit in common use is the **horsepower**, which is defined as 550 foot-pounds per second, and is therefore based on the gravitational system. It is not so large as a kilowatt, for it takes only 746 watts to make a horsepower.

Both the kilowatt and horsepower give rise to two more energy units, the **kilowatt-hour** and the **horsepower-hour**. Since power is work divided by time, power multiplied by time measures work or energy. A kilowatt-hour means the work a kilowatt of power does in an hour, and similarly for the horsepower. Therefore to reduce the former unit to joules, we must multiply a thousand watts by the number of seconds in an hour, or $1000 \times 3600 = 36 \times 10^5$ joules. One horsepower hour = $550 \times 3600 = 198 \times 10^4$ foot-pounds.

64. Potential energy. The ability to do work is always due to one of two different causes, or to both combined. The first of these is the arrangement, or state of the bodies concerned, and is called **potential energy**. The second is due to motion and is called **kinetic energy**. Potential energy is essentially energy of position or condition stored up in the system in a latent form. A watch spring or the weights of a clock after winding, and a charge of gunpowder all possess potential energy. The wound-up watch and clock are clearly cases where the *position* of the bodies represents energy, the first being due to stresses set up in the bent steel, and the second to the tendency of the raised weights to fall again. But gunpowder is an illustration of energy of *state*. Its explosion represents chemical changes in the ingredients due to their previous unstable condition.

65. Calculation of potential energy. The energy latent in explosives can be calculated only from chemical data regarding the heat of combination of their components. But when the energy is due to some arrangement like a coiled spring or the raised weights of a clock, we have only to measure the work that was expended in getting them in that condition. This is very easy when the force is constant, like gravity. Thus a ten-pound weight raised three feet represents thirty foot-pounds of energy or 32×30 foot-pounds. In the c.g.s. system, 5 kg raised a meter has potential energy given by $mgh = 5000 \times 980 \times 100$, or 49×10^7 ergs, or 49 joules.

An important illustration of potential energy is furnished by laying n flat blocks of stone on top of each other so as to form a vertical column. The work done in putting the topmost block in place is the product of its weight mg and the height H through which it is raised. So its potential energy is

$$W = mgH. \quad (1)$$

If the blocks are many and thin, we may say that the height of the column is the distance H through which the topmost block is raised, though really it is less by the thickness of the block. Then the aver-

age height to which any block is raised is $H/2$, and the work done in building the column, or its entire potential energy, is given by

$$W = \frac{nmgH}{2}. \quad (2)$$

This is only approximately true for thin blocks, but in the case of a continuous medium like a standpipe filled with water, where nm is the total mass lifted, the calculation is exact and much used in hydrostatic problems.

66. Potential. If a mass of one gram is raised against the force of gravity from a certain level to a height h centimeters above that level, the work done per gram is gh ergs. This is called the gravitational potential of the point to which the gram was raised, with reference to the original level. This potential then is measured in terms of work per unit mass. Thus if a mass m is lifted to the higher level, the work done is mgh ergs, and the potential is obtained by dividing by m , giving gh ergs per gram as before.

Potential depends upon position in space, and can be calculated for points in gravitating systems, without performing any experiment. Its unit is an erg per gram, and its dimensions are $[FL/M] = [L^2/T^2]$ or $[L^2T^{-2}]$, which is the square of a velocity. If the plane of reference is sea level, then the top of a rock ten meters above the surface of the sea is at a potential of $1000 \times 980 = 98 \times 10^4 \text{ cm}^2/\text{sec}^2$.

In the gravitational system, work is measured by wh , where w is the weight, and the potential wh/m has the numerical value of h because weight is taken numerically equal to the mass. This height h is expressed as a "head" of so many feet in problems in hydraulics using the British units. But head is really neither a height nor a pressure, but a potential, and is measured by the work done against gravity in raising a pound mass from one level to another through a vertical distance h .

There is a great variety of potentials used in physics, and in general, this concept is valuable wherever the potential energy of a system depends upon some measurable kind of work. It is always calculated as the work concerned with getting some unit quantity into the configuration or state considered. *Work per unit quantity*, then, is a broad definition of potential, provided the "quantity" is suitably chosen, and provided the work is measured in terms of the characteristic forces of the system.

67. Kinetic energy. Energy due to motion is called **kinetic energy**, from the Greek word *kinein*, meaning "to move." When a rapidly

moving body such as an automobile strikes an obstacle, it is capable of doing great damage. This damage means *work* in the mechanical sense. Therefore the moving object possesses the *capacity for doing work* upon objects with which it may collide. In other words it possesses *energy* due to motion relative to something else.

It is only through the force of impact with another body that kinetic energy manifests itself, and is able to perform work. Therefore, in a sense, the kinetic energy of a projectile is latent until it strikes the target. It is then able to perform work because of its relative *motion*.

This procedure should be contrasted with potential energy similar to that of a clock weight. It is due to relative *position* and is latent until the weight begins moving.

68. Calculation of kinetic energy. The most logical way of obtaining the amount of energy due to the motion of a body is to calculate the work required to give it that motion. This means measuring the work which is converted into the kinetic energy of a moving body and which is ready to do work when it is stopped. Or, what amounts to the same thing, we may calculate the work required to stop a moving body. If a constant force acts directly upon a free body through a distance s , the work done is $W = Fs$. But $F = ma$, and according to equation (4) of accelerated motion, the velocity acquired from rest is $v^2 = 2as$, or $s = v^2/2a$. Substituting these values, we obtain

$$W = ma \times \frac{v^2}{2a}$$

and

$$W = \frac{1}{2}mv^2, \quad (1)$$

which is the kinetic energy in terms of the mass and velocity of the moving body with reference to other bodies regarded as at rest. Its unit and dimensions are obviously the same as those of potential energy, or

$$[W] = [ML^2T^{-2}].$$

In the gravitational system of units, $F = ma/k$; therefore the kinetic energy is given by

$$W = \frac{mv^2}{2k}. \quad (2)$$

Using the value of k suited to the units in which m and v are measured, as explained in Article 50, we obtain from equation (2) the kinetic energy in the gravitational units: gram-centimeters, kilogram-meters, or foot-pounds.

If m is expressed in slugs, then $mv^2/2$ gives the energy in foot-pounds directly, because expressing the mass in slugs amounts to setting k equal to unity.

69. The conservation of energy. Let us suppose that we give kinetic energy to a body by dropping it from a definite height above a level surface. If the body is highly elastic like a new tennis ball, and the surface hard, it will rebound to a height only a little below where it started. If the experiment were performed in a vacuum it would rise higher still, but in no case would it quite regain its original position. However, we may infer that it would do so if it were perfectly elastic, if the surface were perfectly hard, and if the air resistance were eliminated.

In such an ideal case the potential energy of the ball at the moment it is dropped is progressively converted into kinetic energy as it gains velocity, and when it strikes the surface at zero potential, the energy is all kinetic. If we assume perfect elasticity, there is no lost energy during the impact, and the ball starts upward with the velocity that it has acquired during its fall. Then if there were no air resistance it would rise to the exact height from which it started, when its kinetic energy would be zero, but it would have recovered its original potential energy.

The experiment just described is an illustration of a basic principle of our universe known as the *conservation of energy*. According to this principle, which has always been found true whenever tested by experiment, energy can be neither created nor destroyed, but only transformed from one kind to another. A wound watch spring transforms the work exerted in winding it into potential energy which reappears as the kinetic energy of the moving wheels and in work done against friction. The potential energy of gunpowder reappears mainly as the kinetic energy of the projectile. Even when some energy seems to be lost, it can always be accounted for as heat, light, sound, or some other less obvious manifestation.

According to modern views introduced by relativity, mass itself may be regarded as a form of energy, and when it emits heat and light as the sun is doing, the mass itself decreases to supply this energy of radiation. If this view is accepted, the law of the conservation of energy must be extended to include mass, so that the conservation of the totality of the mass and energy of a system would replace the older statement.

It is a remarkable fact that, in general, potential energy tends to become kinetic, as occurs when rocks roll down a mountain,

while kinetic energy is transformed into potential only under rather limited conditions; for example, when water is being pumped up into a tank. Generally speaking, both forms of energy tend toward motion of some sort; the most common is the motion of the molecules of a body known as heat. Heat is a kind of kinetic energy, and appears to be the ultimate form in which other forms of energy tend to appear. The energy of the stone rolling down the mountain is frittered away as heat along its route and when it hits the bottom. This tendency is called the "degradation" of energy, because as heat, it is not ordinarily available for useful work.

70. Comparison of momentum and kinetic energy. As we have seen, the impulse of a force, or the force multiplied by the *time* during which it acts, measures the momentum mv which the body acted on acquires. But the product of the force and the *distance* through which it acts gives us the kinetic energy $\frac{1}{2}mv^2$. If the momenta of bodies having different masses are equal, as in the case of the gun and its projectile, then the kinetic energies are unequal. The lighter body has a corresponding higher velocity, and since this is squared, it more than compensates for its lesser mass in calculating $\frac{1}{2}mv^2$.

The converse of the above is also true. If two bodies of different masses have the same kinetic energy, their momenta are not equal. The larger mass has the greater momentum in spite of its smaller velocity.

In the case of a moving mass striking a stationary object without rebounding, as when the descending block of a pile driver strikes the pile, the resulting motion of the latter depends upon the momentum of the block. Therefore, if a fixed amount of kinetic energy is available, it is more effective to use a heavy mass moving at a relatively slow speed than a lighter mass moving at a high speed. The momentum of the slow-moving heavy mass after falling a short distance is greater than that of a small rapidly moving mass which has been lifted higher by the expenditure of the same amount of energy.

If, however, a comparison is made between two masses having equal momenta, the light one with its higher velocity has more energy, and would, for instance, develop more heat in a target which completely stops it, than would the heavy slow-moving projectile.

Another important difference between kinetic energy and momentum is that, after a collision, the total momentum is unchanged as such, while the kinetic energy reappears in various forms. Thus if a rifle bullet is stopped by a wooden block suspended like a pendulum, the block (with the embedded bullet) acquires the entire momentum of

the bullet. If the velocity given to the block is measured, the bullet's velocity may be calculated from the equation $(M + m)V = mv$. But the kinetic energy acquired by the swinging block is always much less than the original $\frac{1}{2}mv^2$, because a great portion of it is transformed into heat.

71. Calculations involving kinetic energy. If it is desired to find the *distance* through which a force must act on a mass m to produce a velocity v , or what velocity is acquired when a force acts upon the mass through a distance s , then the equation to be used is $Fs = \frac{1}{2}mv^2$, just as $Ft = mv$ answers the same sort of question when the *time* of action is involved. Suppose a man pulls horizontally on a sled whose mass is 10 kg and which rests upon smooth ice. If he uses a force of a million dynes for a distance of a meter, the velocity acquired is given by $10^6 \times 100 = \frac{1}{2}10^4 \times v^2$, whence $v^2 = 2 \times 10^4$, and $v = 100\sqrt{2}$ cm/sec. Conversely he could have stopped the same mass moving at this velocity of $100\sqrt{2}$ cm/sec. in the distance of a meter by applying a force of a million dynes. In the gravitational system such problems require the introduction of k (unless the mass is expressed in slugs), as has already been pointed out. Then if the force is 40 lb. and acts on a mass of 64 lb. for a distance of 10 ft., the resulting velocity is given by $40 \times 10 = \frac{1}{2}(64v^2/32)$. Hence $v^2 = 400$, and $v = 20$ ft./sec. The distance a moving train travels after the brakes are applied to produce a retarding force, is a problem of practical importance to be solved in this manner. We may also calculate the distance a train goes before a given pull by the locomotive has resulted in a required velocity.

72. Equilibrium. The conditions under which a number of actions may be so balanced that no motion results have been explained.

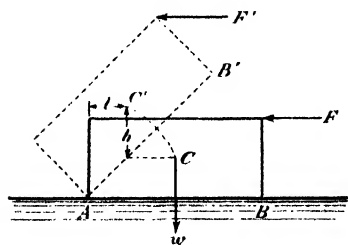


Fig. 44.

But we shall now examine, from the standpoint of energy, what happens when an unbalanced action disturbs the existing equilibrium.

There are three cases: when the body gains potential energy under the action of the disturbing force, when its potential energy remains constant, and when it decreases.

To illustrate the first case, suppose

a force F acts on the block shown in Fig. 44 so as to produce a torque about the corner A which is hinged to the plane AB . As the force tilts the body up, the center of gravity C rises through a vertical

distance h , and the potential energy of the block rises in proportion. But the force w acting on the lever arm l produces a restoring torque tending to lower the potential energy and create kinetic energy. If now left to itself the block will fall back into its original position, where it is said to be in **stable equilibrium**. This condition means that every possible motion which alters the level of the center of gravity must *begin* by raising it, and thus increasing the potential energy.

If F continues to act, causing further rotation about A , the situation shown in Fig. 45 is reached. Here C' is at the highest point of the circular arc it is tracing and directly over the point of support. The force w has no restoring moment and the block just balances. In this condition any motion in either direction will tend to lower C' , and the block will fall in that direction. This is known as **unstable equilibrium**. In a still broader aspect, the potential energy of a system, due either to position or state, is unstable when a very small action precipitates a change to kinetic energy. The spark which ignites a charge of gunpowder is an example of a small action producing a violent transformation from unstable potential energy into the kinetic energy of the explosion.

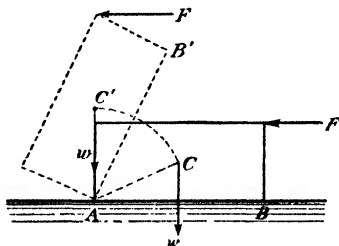


Fig. 45.

In addition to stable and unstable equilibrium, there is a third type known as **neutral equilibrium**. This is represented by a ball free to roll on a horizontal plane. Unless raised bodily from the plane, no force tends to raise or lower its center of gravity, and it remains at rest indifferently in all positions, because no motion on the plane alters its potential energy.

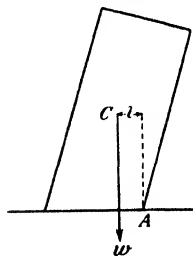


Fig. 46.

The Leaning Tower of Pisa represented in Fig. 46 is an interesting illustration of stable equilibrium. The vector w representing the tower's weight falls inside the base. This creates a restoring torque wl tending to lower the center of gravity if an overturning force should succeed in raising it slightly. The tower is therefore in stable equilibrium unless C should be brought directly above A . Then its equilibrium would be unstable and a slight excess push would wreck it.

73. Friction. When two bodies are in contact and one moves or tends to move over the other, there is always a force known as **friction** which opposes the motion. There are two principal kinds to be considered: sliding friction, and rolling friction, and each may be divided into static and kinetic varieties. Sliding friction is represented by countless phenomena. It makes walking possible, holds the nails in the walls of our houses, slows down our automobiles when the brakes are applied, and makes possible all manner of earth works, such as railroad embankments. In all such cases friction is absolutely essential. But very often, on the contrary, it is most undesirable and is eliminated as far as possible by the use of lubricants, as in the moving parts of machines.

Since the force of friction opposes the motion of a sliding body, its value before motion begins is, of course, always equal and opposite to any force which acts in the direction of the sliding motion. The force of friction then goes on increasing, as the applied force increases, from zero up to a maximum just before motion begins, and then falls off slightly when the body is actually moving. The maximum value thus obtained is the static form of sliding friction, while the smaller value after motion takes place is the kinetic form. Static sliding friction is greater than the kinetic form because the slight roughnesses of the two bodies seem to settle into more intimate contact when at rest. After motion begins, it is found that just as much force is needed to pull a body rapidly as slowly over a surface, unless the speed is very high. This means that within wide limits sliding friction is independent of the relative speed of the surfaces in contact.

But speed is not the only factor to be considered. The question naturally arises as to how friction is affected by the normal force with which two bodies are pressed together. If a wooden slab sliding on a level board is loaded with increasing weights, it is found that the forces required to start it and then to keep it moving at an even pace vary directly as the total weight.

Finally we may wish to know how friction is affected by varying the area of contact without changing the force. If the wooden slab is placed edgewise instead of flat, but loaded as before, it is found that the pull needed to overcome the opposing force of friction is unaltered. This seems rather surprising, but is easily explained by considering that decreasing the area where friction occurs is offset by the increased force per unit area over that surface. That is, the *pressure* increases as the area diminishes, if the total normal force remains constant. If

the pressure instead of the force is kept constant, then the force of friction varies directly with the area.

74. The coefficient of friction. The experimental facts just described may be formulated as the so-called "laws of friction." Like other laws, these are only concise statements which summarize observed phenomena in a convenient form. The "laws" of friction are: (1) That within limits, the force of kinetic friction is independent of the velocity; (2) That both static and kinetic friction vary with the normal force pressing two surfaces together; (3) That the force of friction is independent of the areas in contact. An analytical expression of these facts is given by the simple relation

$$f = \mu N,$$

where N is the normal force and μ is the constant of proportionality known as the coefficient of friction. This important quantity differs with different surfaces, depending upon their smoothness and material. We may then formulate a fourth law, that when the normal force between two bodies is constant, the force of friction depends only upon the nature of the surfaces in contact.

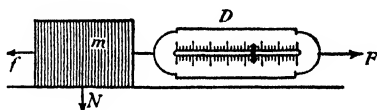


Fig. 47.

To determine the value of the coefficient μ for a given pair of surfaces, we may perform a simple experiment illustrated in Fig. 47. A block of mass m is dragged horizontally at constant velocity by a force F over a level surface, with a dynamometer D arranged to record the force of the pull. Then $N = mg$, and $F = \mu mg$. This is true only when there is no acceleration, so that kinetic friction is the only force opposing the motion.

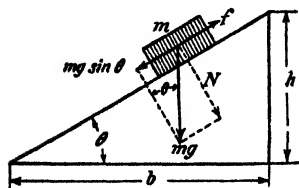


Fig. 48.

If there is a positive acceleration, the dynamometer will record a greater force, equal to $\mu mg + ma$. If the pull is less than sufficient to move the mass m , the force of friction is less than μmg , for it is always equal and opposite to the applied force when there is no acceleration.

A convenient way of measuring the coefficient of static friction is to tilt the plane up until m starts sliding. In this case the component of gravity acting along the plane is $mg \sin \theta$, as in Fig. 48. It is equal and opposite to the force of friction. But the latter is due

to N , the normal component of mg , and not to the total weight as above. Since $N = mg \cos \theta$, the equation of equilibrium becomes

$$mg \sin \theta = \mu mg \cos \theta.$$

Hence

$$\mu = \tan \theta = \frac{h}{b},$$

so that μ is calculated from easily measured quantities. The angle θ , when the block just starts to move, is known as the "limiting angle of repose." If a pile of stones or an earth embankment is to be stable, it must lie at an angle smaller than θ . The "talus" of rocks at the foot of a mountain cliff automatically assumes the angle of repose, as it steadily grows owing to the cliff's disintegration.

In order to find the coefficient of friction during motion, the device of the plane may again be used, provided that after the block starts, θ is slightly decreased to the value which is just sufficient to keep the block moving at a steady speed. This gives the value of the coefficient of kinetic friction.

75. Rolling friction. When one body rolls upon another there is theoretically no slip between them. And if both are perfectly hard,

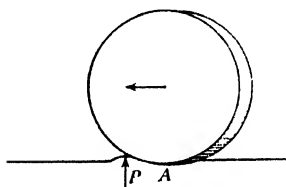


Fig. 49.

there is no *surface* of contact, but only a point where a sphere would rest, or a line when a disc rolls upon a plane. If the supporting surface is not perfectly hard, the disc causes a slight depression at A which results in the formation of a ridge at P in front of it, as shown in Fig. 49.

This shifts the line of contact forward and upward, so that the disc is being continually pulled up a minute hill. The force required to do this may be calculated by the principle of moments.

In this case, as in sliding friction, static friction is greater than kinetic friction, so that quite apart from the force required to accelerate a train from rest, the locomotive has to exert a greater force to overcome static rolling friction than is needed when the train is in steady motion. In consequence of this the train may be unable to start, although when once in motion the locomotive will be equal to the task.

76. Problems involving friction. The problems of bodies sliding on horizontal or inclined planes become somewhat more difficult when the effect of friction is introduced, although the same formulæ as previously derived are used, but with the addition of a new force.

The block resting upon a horizontal plane, shown in Fig. 50, requires a force equal to $f + ma$ to start it moving, where f is the force of friction equal to μmg (when F is horizontal), and ma is the inertia reaction; therefore the acting force is $F = \mu mg + ma$. If F , however, acts at an angle θ , then both terms are modified. The horizontal component of F , $F \cos \theta$, is its effective value, and the normal force N is no longer only mg , but $mg + F \sin \theta$, if F is a push; or $mg - F \sin \theta$, if it is a pull, since F has a vertical component which changes N and the force of friction at the same time. Therefore the equation of actions and reactions becomes

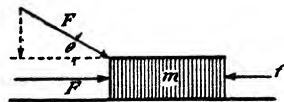


Fig. 50.

$$F \cos \theta - \mu(mg \pm F \sin \theta) - ma = 0. \quad (1)$$

This makes it possible to calculate a if the other quantities are known, and then we may use any of the equations of uniformly accelerated motion, as in the case of freely falling bodies.

On an inclined plane, as in Fig. 51, if the angle θ is greater than the angle of repose, the body slides downward with accelerated motion. The various actions are F_g (the component of mg acting along the plane), the opposing force of friction $f = \mu N$, and the inertia reaction ma . Then

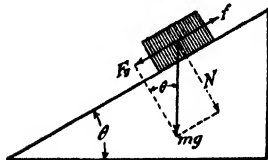


Fig. 51.

$$mg \sin \theta - \mu N - ma = 0,$$

and

$$mg \sin \theta - \mu mg \cos \theta - ma = 0. \quad (2)$$

Then m may be cancelled, and as g , μ and θ are constant, a is also, and may be calculated and used in the various equations of uniformly accelerated motion.

77. Simple machines. The six devices known as simple machines are the lever, the inclined plane, the wheel and axle, the screw, the wedge, and the block and tackle. Of these the first two are radically different from each other, but the wheel and axle, and the block and tackle may be treated as levers, while the screw and wedge are derived from the inclined plane.

In general, a machine increases the force applied with a loss in the amount of motion, but it is sometimes used to increase the motion with a corresponding decrease in the force overcome. This follows from the law of the conservation of energy, according to which the work done on the machine must be equal to the work done by it plus

the losses which appear as heat. We may express this principle symbolically by

$$W_i = W_o + S.$$

$$\therefore F_i l_i = F_o l_o + S,$$

where W_i is the energy input, W_o is the useful output, S represents the losses, and F and l are the corresponding forces and the distances through which they act.

78. The lever. A rigid bar free to move about an axis or *fulcrum*, constitutes a lever. In Fig. 52 if the forces F_a , F_b , and R , as well as their moments about the fulcrum O , satisfy the requirements of

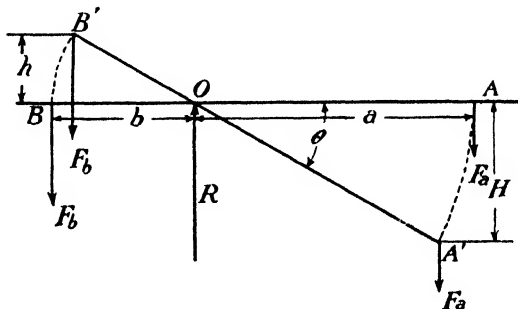


Fig. 52.

equilibrium ($\Sigma F = 0$, $\Sigma Fl = 0$), then there is no motion. If lever arm a is greater than b , then $F_b > F_a$ according to the proportion $F_a : b :: F_b : a$. Therefore a small force, acting at A , balances a large force at B . But a machine is of no value unless it does work, and this implies motion. Suppose then that F_a acts through a vertical distance H , causing the bar to rotate through an angle θ , thus raising B vertically through a distance h . Then the work F_a does is $F_a H$, and that done against the opposing moment of the force F_b is $F_b h$. If friction is neglected, these are equal and

$$F_a H = F_b h.$$

But by simple geometry

$$\frac{H}{a} = \frac{h}{b}.$$

Hence

$$\frac{F_a}{b} = \frac{F_b}{a},$$

just as before. That this relation holds for any angle may also be shown by moments without equating the work done, for $F_a a \cos \theta =$

$F_b b \cos \theta$ when the two torques at any angle θ are equal. Therefore $F_a : b :: F_b : a$.

If this is true in any chosen position, then it must be continually true as the bar rotates at a constant rate. It should however be noted that if the object is to raise B , this is best accomplished when $\theta = 0$, for then the vertical motion is greatest. When $\theta = 90^\circ$ there is no vertical motion at all. In case the forces are normal to the bar in all positions instead of remaining vertical as just assumed, then the angle θ has no significance.

The ratio of the two forces $F_b : F_a$ is known as the **mechanical advantage** of the machine. In the lever it equals the ratio of the arms, $a : b$. But in any machine it may be expressed as the ratio of the force exerted by the machine to the force applied.

There are three possible arrangements of the forces concerned in the action of the lever. The one just mentioned is a "lever of the first class" with the fulcrum between the two forces. It is the most common kind, and is seen

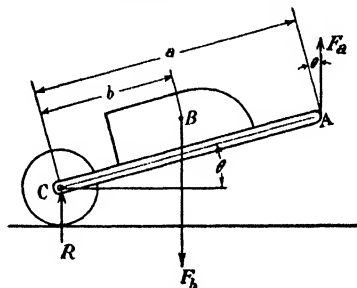


Fig. 53.

in the crowbar as it is ordinarily used in prying up weights, in a pair of pliers where the jaws are short and the grip long, and in many other familiar devices.

"Levers of the second class" have the fulcrum outside the two forces. A wheelbarrow acts on this principle. As seen in Fig. 53, $F_a a \cos \theta = F_b b \cos \theta$, and a relatively small force applied at A exerts a large force upon the load whose center of mass is at B .

If A and B are reversed, so that the active force is at B and the load at A , the machine acts at a disadvantage. The force exerted is less than that applied, although the load is moved through a correspondingly greater distance. This is the "lever of the third class," and is well represented by the human forearm, as shown in Fig. 54. The elbow joint C is the fulcrum, and the load at B acting on a long lever arm is supported by the muscle at A acting upon a short lever arm.

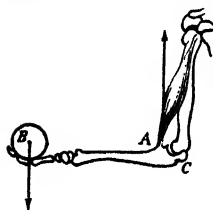


Fig. 54.

The wheel and axle is a modified lever of the first class and capable of continuous motion. Here the force $m_1 g$ acting upon a lever arm

r_1 (the wheel's radius, Fig. 55) may more than counterbalance the opposing torque due to the larger mass m_2 which acts upon the shorter lever arm r_2 , the radius of the "axle."

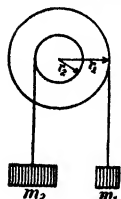


Fig. 55.

79. The inclined plane. This device, regarded as a machine, enables a relatively small force to raise a large mass through a vertical height which might be impossible by a vertical push or pull. Thus, if friction is neglected, the component of gravity acting down the plane, or $mg \sin \theta$, is balanced by the force F , as shown in Fig. 56. If θ is small, mg may be vastly greater than F , and we obtain the advantage expressed by

$$\frac{mg}{F} = \frac{1}{\sin \theta} = \frac{l}{h}, \quad (1)$$

which may be made as large as desired by having h small compared to l .

If, however, the force acts horizontally, the effective component $F \cos \theta$ replaces F in (1), and we obtain $F \cos \theta = mg \sin \theta$. The mechanical advantage is therefore

$$\frac{mg}{F} = \frac{1}{\tan \theta} = \frac{b}{h}. \quad (2)$$

For small angles, b equals l very nearly and there is no appreciable loss of advantage in using a horizontal force. But as θ increases, l becomes steadily larger than b , so that l/h becomes steadily larger than b/h . This means that a horizontal force, as compared to one along the plane, acts at an increasing disadvantage. At 45° , $b = h$, the advantage for a horizontal force is unity, and the plane is no longer of any value. Beyond 45° the plane is a disadvantage and when $\theta = 90^\circ$, motion due to a horizontal force is impossible.

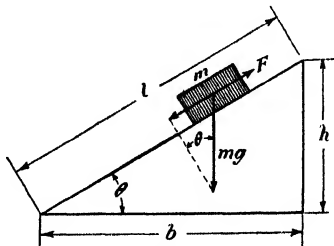


Fig. 56.

The foregoing results, where no friction is involved, can also be obtained by the method of equating work input to output. When the load reaches the top of the incline, the work done is the weight multiplied by the vertical height through which it has been raised, or mgh . This is accomplished by the application of a force acting through a distance l in case (1) and b in case (2). Then equating input to output, we have Fl (or Fb) = mgh , as shown above.

In the case of the wedge, the inclined plane is forced between two objects instead of having one of the objects moved along it. If the angle of the wedge is small, an enormous force is exerted. In Fig. 57 the block A is fixed, and the wedge W is pushed under B by a force F . This is the same as the horizontal force required to push B up the plane with W stationary, or $F = mg \tan \theta$, in equation (2) above.

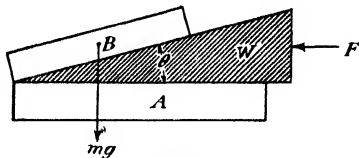


Fig. 57.

80. The screw. This is another form of inclined plane, as used in the "jack" to lift heavy objects. The *thread* is a helical plane wound around a cylinder at an angle which corresponds to the slope of the ordinary inclined plane. If there are, for instance, two threads to

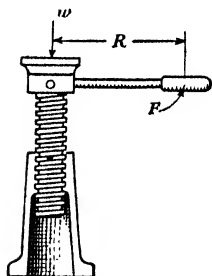


Fig. 58.

the centimeter, it will take two turns of the screw jack to raise an object that distance. This is usually accomplished by some sort of crank, or lever, or gear. Suppose the force is applied normally to the end of a crank of radius R , as suggested in Fig. 58; then if the screw has a pitch p , which is the vertical distance between threads, the work done on the crank in one turn is $2\pi RF$, and this accomplishes $w p$ units of work in lifting the load. These two amounts are equal

if we ignore the very large item of friction in this particular machine. Then $2\pi RF = wp$, and the mechanical advantage becomes $w/F = 2\pi R/p$. This ratio is seen to increase both with the length of the crank and the number of threads to the centimeter, or the smallness of the "pitch."

81. The block and tackle. A rope running through a system of "pulleys" may produce very great forces from relatively small ones. Like the wheel and the axle, pulleys may be treated by equating moments, but they are more simply explained as follows: In the simplest case with a block of one *sheave* (the grooved wheel over which the rope passes), we

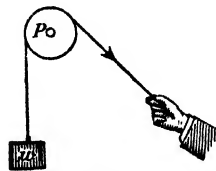


Fig. 59.

obtain only a change of direction of the force, as shown in Fig. 59, and the weight moves upward as much as the hand moves in the direction of the arrow. If it is desired to increase the mechanical advantage, there must be at least one movable block. This may be

accomplished with a single turn, as shown in Fig. 60 (a), where the end of the rope is fixed at P , and the applied force acts upward. Here the weight w moves only half as fast as the free end of the

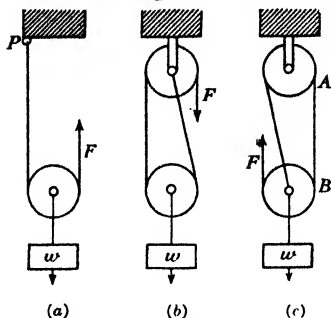


Fig. 60.

rope, so that $Fs = \frac{1}{2}ws$, where s is the distance through which F acts, and the advantage is 2:1. If the rope passes once more around the fixed block, as in Fig. 60 (b), the direction of F is reversed with no further gain in mechanical advantage, though the arrangement is more convenient. In Fig. 60 (c), the rope starts at the movable block and a force applied upward raises w one third of the distance through which it acts. Thus

the mechanical advantage, if friction is always ignored, is 3:1. This could again be reversed as in (b), without gain in mechanical advantage, by another turn around A equipped with two wheels or "sheaves." A second turn around B would result in an advantage of 5:1 and so on indefinitely with every strand supporting the movable block adding a unit to the result. It is well to note that the mechanical advantages are always even numbers when the rope starts at the fixed pulley, otherwise odd.

In the pulley systems just described there is only one cord, but there are other systems using two or more distinct cords. Thus the cord supporting the weight in Fig. 60 (a) might pass around a second movable block and then upward to a rigid support, as shown in Fig. 61. The first block would be pulled upward with a force $2F$. The second block would be pulled upward by a force equal to $4F$, or 2^2F . Then a third cord might originate at the second block, and passing around a third block, would lift it with a force of $8F$, or 2^3F . Thus the mechanical advantage of such a system would vary as 2^n , where n is the number of blocks. The tension on the cords shown in Fig. 60 is of course equal to F , but when several are used as described above, the succeeding cords experience tensions of $2F$, $4F$, $8F$, and so forth, so that they should be of correspondingly increasing strength. There are still other and more complicated systems, but these are not often used.

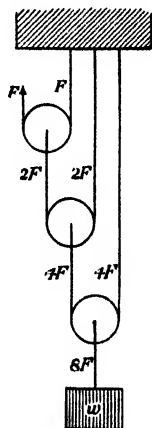


Fig. 61.

82. Efficiency of machines. In all machines, friction must be considered as of more or less importance. In a lever acting on a very sharp and hard fulcrum, there is almost none. But in such machines as the screw, and the block and tackle, there is a great deal of friction. Its effect is to lower the mechanical advantage from its ideal value, for the active force must do work against friction as well as furnish the energy required by the load. Thus the useful energy output is less than that which we have calculated in the preceding ideal cases, because part of the input has been wasted in overcoming friction and frittered away as heat.

In general the term **efficiency** is applied to the performance of any device which is used to transform energy. It is the ratio of the total work produced to that which was transformed in producing it. Or more briefly, it is **work output divided by work input**. If these are

equal the efficiency is unity, or 100 per cent, and this is the case in the ideal machines we have been considering. But no such device is devoid of friction or some other kind of loss, so that the efficiency is always less than the ideal 100 per cent. We have therefore no right to equate output to input, but should write $W_i = W_o + S$, where W_i is the energy input, W_o the useful output, and S the losses. The efficiency then is given by $W_o:W_i$, or $W_o:(W_o + S)$. It falls increasingly below 1 as the quantity S becomes more and more important.

83. The balance. Although not a machine in the strict sense of the word, this very important device for comparing masses, or

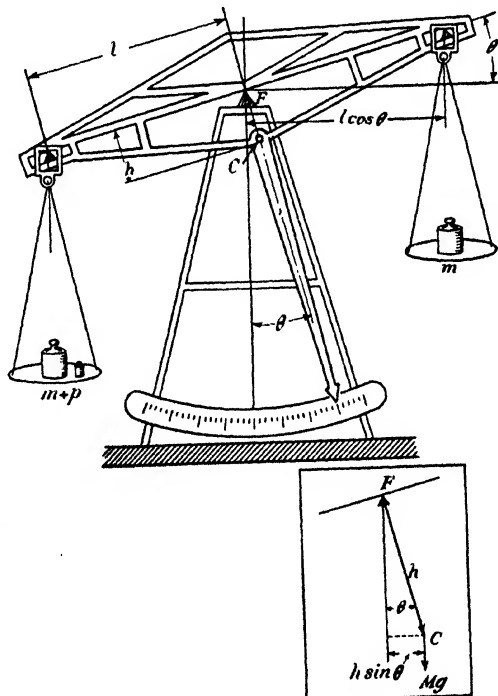


Fig. 62.

"weighing" bodies, will be considered here, since it depends upon the same principles as the lever. If the arms shown in Fig. 62 are of equal length l , then the moments due to the two equal masses m in the pans are equal, $mgl = mgl$, and the beam remains horizontal. But if an additional small mass p is placed in one of the pans, a new moment, pgl , disturbs the equilibrium and must be counteracted by an equal and opposite torque if equilibrium is to be restored. This occurs after the beam has assumed a new position at an angle θ with the horizontal. Let M be the mass of the beam, supposed concentrated at C . Let C be h centimeters below the fulcrum on a line normal to the beam. Then there is a restoring torque equal to $Mgh \sin \theta$. But the torque due to p is now no longer pgl , but $pgl \cos \theta$. Equating these values, we obtain $pgl \cos \theta = Mgh \sin \theta$,

$$\text{or} \quad \frac{\tan \theta}{p} = \frac{l}{Mh}.$$

For small angles, $\tan \theta = \theta$, nearly, and the ratio $\theta:p$ (with p usually given in milligrams) is a measure of the sensitivity of the balance. This may be defined as the angle (or number of scale divisions) swept over by the pointer per milligram. The sensitivity is clearly increased by using long arms, a small mass M (light construction), and by bringing the center of mass near the fulcrum (h small). Each of these conditions has serious drawbacks when carried too far, and the design of a delicate balance consists in a judicious adjustment between advantages and disadvantages.

SUPPLEMENTARY READING

H. A. Erickson, *Elements of Mechanics* (Chap. 9), McGraw-Hill, 1927.
J. S. Ames, *The Constitution of Matter* (Chap. 1), Houghton Mifflin, 1913.

PROBLEMS

1. How much work is required to build a column 12 ft. high of 6 granite blocks each 2 ft. thick and weighing 500 lb.? *Ans.* 15,000 ft.-lb.
2. How much power is required to pump water at the rate of 90 l a minute to a height of 20 m? *Ans.* 294 watts.
3. How fast can a derrick lift a block of marble weighing 1100 lb. if 4 hp. is applied to the tackle? *Ans.* 2 ft./sec.
4. What is the potential energy of the water which fills a cubical tank whose edge is 10 ft. and whose base is 20 ft. above the ground? (A cubic foot of water weighs 62.4 lb. approximately.) *Ans.* 1.56×10^6 ft.-lb.

5. A piston sliding with negligible friction in a vertical cylinder 1 ft. in diameter is loaded to weigh 150 lb. Water is forced under it so as to raise it 4 ft. What is the potential energy of the system? *Ans.* 992 ft.-lb.

6. Calculate the gravitational potential with respect to sea level at an altitude of 50 m. *Ans.* 4.9×10^6 ergs/g.

7. What is the kinetic energy of a mass of 50 g moving with a speed of 800 cm/sec.? *Ans.* 1.6 joules.

8. What is the kinetic energy in ft.-lb. of a bullet weighing 8 oz. and traveling at a speed of 1800 ft./sec.? *Ans.* 25,155 ft.-lb.

9. Calculate the kinetic energy acquired by a mass of 10 kg after falling for 2 sec. from rest. *Ans.* 1921 joules.

10. What are the momenta and kinetic energies of a gun, free to move backward, and its projectile, if the gun weighs 10 kg, the projectile 80 g, and its muzzle velocity is 400 m/sec.? *Ans.* Momentum of each is 3.2×10^6 g-cm/sec.; k.e. of the projectile is 6400 joules, of the gun 51.2 joules.

11. How far must a man pull a garden roller weighing 200 kg in order to get it going at 3 m per second, if he exerts a uniform pull of 60 kg? *Ans.* 1.53 m.

12. An automobile weighing 3000 lb. and going at 40 miles per hour is stopped by the brakes in 50 ft. What is the retarding force in pounds? *Ans.* 3207 lb.

13. A locomotive exerts a "draw-bar pull" of 20 tons on a stationary train weighing 800 tons. What is its velocity at the end of a quarter of a mile? *Ans.* 46.1 ft./sec.

14. What force is required to pull a block weighing 3 kg along a horizontal surface at constant speed, if the coefficient of friction is 0.27? *Ans.* 0.81 kg.

15. What constant force must be applied to the block in Problem 14 in order to get it going at 2 m per second from rest in 5 sec.? *Ans.* 932 g.

16. A steady force of 10 lb. applied as in Fig. 50 at an angle of 30° to the horizontal acts on a block weighing 15 lb. which rests on a horizontal surface. In 2 sec. the block is moving with a speed of 18 ft./sec. What is the value of μ ? *Ans.* 0.223.

17. What is the acceleration of a block sliding down a 30° slope, when the coefficient of friction is 0.25? *Ans.* 9.14 ft./sec².

18. What force acting along the plane is needed just to prevent the motion of a 4 lb. block as in Problem 17? What force is needed to cause the block just to move up the plane? *Ans.* 36.56 poundals; 92.3 poundals.

19. What horsepower is exerted in pulling a 200 lb. log up a 30° slope at the rate of 12 ft./sec., when the coefficient of friction is 0.3? *Ans.* 3.3 hp.

20. Find the time required for a block to slide to the bottom of a plane 640 cm long if the incline is 30° , and the coefficient of friction is 0.2. *Ans.* 2 sec.

21. If the block weighs 50 g, calculate the initial potential and the final kinetic energies. Account for the difference. *Ans.* 1.57 joules; 1.02 joules; lost energy due to friction appears at heat.

22. A body starting from rest slides down an inclined plane whose slope is 30° . The coefficient of friction is 0.2. What is its speed after sliding 76.5 ft.? (In this problem and the next take g as 32 ft./sec².) *Ans.* 40 ft./sec.

23. An object of any weight rests upon a 30° slope. It is given a shove downgrade so that its initial speed is 12.8 ft./sec. The coefficient of friction is 0.6. How long will it slide before it is stopped by friction? *Ans.* 20 sec.

24. What is the mechanical advantage of a 6 ft. crowbar when the fulcrum is 4 in. from the weight? What is the advantage of the wheelbarrow in Fig. 53, when $a = 6$ ft. and $b = 2$ ft.? *Ans.* 17; 3.

25. What is the mechanical advantage of an inclined plane used as a machine, when $\theta = 30^\circ$, and the force acts horizontally? When it acts along the plane? *Ans.* $\sqrt{3}$; 2.

26. A screw jack has a "pitch" of 0.5 inch. What weight will it lift (neglecting friction) when a force of 20 lb. is applied at a point on the arm 18 in. from the axis? *Ans.* 4521.6 lb.

27. In Fig. 55 the axle has a radius of 2 in. and the wheel has a radius of 18 in. What is the mass of m_1 which just balances 36 lb. represented by m_2 ? *Ans.* 4 lb.

28. What is the efficiency of an engine which delivers 30 hp. for 8 hr. with a consumption of 19.8×10^6 ft.-lb. of energy? *Ans.* 24 per cent.

* 29. If the diameter of the screw in Problem 26 is 2 in., and the coefficient of friction when it is greased is 0.08, what force must be applied at the end of the arm to overcome friction due to a two-ton weight supported by the screw? How much work is done in raising this weight 1 ft.? What is the efficiency? *Ans.* 17.8 lb., 4 ft.-tons, 50 per cent.

* 30. A "stone boat" is drawn up a 20° hill by a rope parallel to the slope. A force of 861.5 lb. is needed to keep it moving at a steady rate, while it takes only 40.7 lb. to pull it down. What are the coefficient of friction and the efficiency of the "machine"? *Ans.* 0.4; 47.6 per cent.

CHAPTER 6

Motion in a Circle

84. Angular velocity and acceleration. When a wheel rotates about its axis we usually describe its velocity as so many revolutions per minute or per second. But a revolution may be measured in angles, and as the unit angle is the radian, we may use radians per second as a measure of angular velocity. As the wheel turns through 2π radians in a revolution, the angular velocity in radians per second, denoted by ω (Greek *omega*), is given by

$$\omega = 2\pi n, \quad (1)$$

where n is the angular velocity in revolutions per second. This measure n is also called the *frequency* of rotation, and the time of one revolution is called the *period*, denoted by T . Then the relation $n = 1/T$ states the obvious fact that if one revolution takes T seconds, the wheel will execute $1/T$ revolution per second, and (1) may be written as $\omega = 2\pi/T$.

Angular velocity is the time rate of angular motion, and when it is constant may be described by $\omega = \theta/t$, just as constant linear velocity is defined by $v = s/t$. In circular measure the angle is the ratio of the arc s to the radius r , or $\theta = s/r$; therefore

$$\omega = \frac{s}{rt}. \quad (2)$$

But $s/t = v$, the linear velocity of a point distant r from the axis; hence

$$\omega = \frac{v}{r}, \text{ or } v = \omega r. \quad (3)$$

Equations (3) are very useful in connecting the linear velocity of a point in a rotating body with the angular velocity of the body as a whole.

Angular acceleration, denoted by α (Greek *alpha*), is the time rate of change of angular velocity, and if the rate is constant, may be calculated from

$$\alpha = \frac{\omega_2 - \omega_1}{t}. \quad (4)$$

In order to connect angular with linear acceleration, we may substitute v/r for ω in (4) which becomes $\alpha = (v_2 - v_1)/rt$. But $(v_2 - v_1)/t$ is the linear acceleration of a point distant r from the axis; therefore

$$\alpha = \frac{a}{r}, \text{ or } a = \alpha r. \quad (5)$$

If velocity and acceleration are not constant, they are defined by $\omega = d\theta/dt$ and $\alpha = d\omega/dt$ respectively.

The dimensions of angular velocity are those of linear velocity divided by a length; hence

$$[\omega] = \left[\frac{L}{T} \times \frac{1}{L} \right] = [T^{-1}],$$

and since $\alpha = a/r$, its dimensions are given by

$$[\alpha] = \left[LT^{-2} \times \frac{1}{L} \right] = [T^{-2}].$$

85. Radial acceleration in circular motion. Constant velocity means constant speed in a straight line. The velocity may then change in two ways. Either the speed or the direction may change,

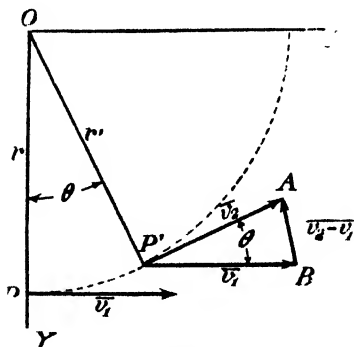


Fig. 63.

and the time rate of either kind is its acceleration. A particle moving with constant speed in a circular path is an illustration of this second kind of acceleration, which in this case is always directed toward the center. In Fig. 63 the velocity of the particle at P is v_1 , but as it rotates counterclockwise about O to P' , v_1 changes to v_2 , and its radius vector r has described the angle θ . Now v_1 and v_2 are both tangent to the circle and normal to r and r'

respectively, and $P'B$ is drawn parallel and equal to v_1 . Therefore, since their sides are mutually perpendicular, the angles $AP'B$ and POP' are equal. The vector BA is the change in the velocity $\overline{v_2 - v_1}$, which added vectorially to v_1 gives us v_2 . So if the angle θ is vanishingly small, or $d\theta$, the corresponding change in velocity AB becomes dv and may be regarded as the arc of a circle drawn with P' as its center and having a radius v . But $d\theta$ is measured by the ratio of the arc dv to the radius v ; therefore $dv = vd\theta$.

The acceleration is the time rate of change of v , which for infinitesimal changes is written $a_r = dv/dt$. If $v d\theta$ is substituted for dv , the radial acceleration is given by

$$a_r = v \frac{d\theta}{dt}.$$

But $d\theta/dt$ is the angular velocity ω , and $v = \omega r$; hence

$$a_r = \omega^2 r. \quad (1)$$

This equation is so important that it will be well to show the various forms in which a_r can be expressed. Thus since $\omega = v/r$, we may substitute for ω in (1) and obtain

$$a_r = \frac{v^2}{r}. \quad (2)$$

Also, since the linear velocity equals the circumference divided by the period T , $v = 2\pi r/T$, and

$$a_r = \frac{4\pi^2 r}{T^2}. \quad (3)$$

Further, substituting n for $1/T$, we obtain

$$a_r = 4\pi^2 n^2 r. \quad (4)$$

To recapitulate,

$$a_r = \omega^2 r = \frac{v^2}{r} = \frac{4\pi^2 r}{T^2} = 4\pi^2 n^2 r. \quad (5)$$

86. Central forces. If you whirl a stone around at the end of a string, it pulls your hand about in the constantly changing direction of the string. If this device is an old-fashioned sling, when you let the stone fly, it goes in a direction at right angles to the string. This may seem strange at first, for its pull was along the string. But there is a simple explanation. All the time you are whirling the stone, it wants to travel in a straight line tangent at any instant to its circular path. But you are forcing it to move in a circle and exerting a *centripetal* (center-seeking) force upon it. The stone reacts with an equal and opposite force commonly called *centrifugal* (center-fleeing). This is really a misnomer because the stone does not tend to flee from the center at all, but to go straight on in the direction it is moving in at any instant, that is, tangentially. So it is better to call the force it seems to exert the *centrifugal reaction*. This reaction is called into being by the force you exert in compelling it to revolve around your hand as a center.

We can understand why the force needed to hold a body in a circular orbit is radial, because we have shown that the body is being constantly accelerated toward the center, and the force causing this change of motion must act in the direction of that change, as was stated by Newton. The reaction to this force is the same as the other reactions we have considered. That is, it is equal and opposite to the action and so is directed away from the center.

87. Calculation of centrifugal reaction. As in linear motion, where $F = ma$, so in rotations, $F_r = ma_r$. The radial acceleration a_r may be calculated from any of the four expressions derived in Article 85. Thus we obtain

$$F_r = m\omega^2 r, \quad (1)$$

$$F_r = \frac{mv^2}{r}, \quad (2)$$

and
$$F_r = 4\pi^2 n^2 m r. \quad (3)$$

Solving (3) for n we obtain a useful expression for the frequency

$$n = \frac{1}{2\pi} \sqrt{\frac{F_r}{mr}}. \quad (4)$$

Equation (3) tells us that the centrifugal reaction of a mass m revolving in a circular orbit varies directly as the square of the frequency, as the mass, and as the radius of the orbit. These facts are most important in the design of flywheels, dynamo armatures, and other devices which rotate at a high speed. If the spokes of a wheel are too light, they may be unable to supply the force needed to hold the rim to its circular path and the wheel "bursts," often with disastrous results.

88. Centripetal forces. These are forces acting toward the center which make the centrifugal reaction possible. They may be due to a variety of causes, the simplest being the cohesive forces between the molecules of the spokes of a wheel or the string used in whirling a stone. This force can meet any value required of it up to the breaking point, and it is quite independent of the radius. A helical spring, however, exerts a force which varies with its extension and it is easy to show that within reasonable limits only one velocity is possible when a stone is whirled about at the end of an ideal spring.

A very important case occurs when the centripetal force F_c varies inversely as the square of the radius. This is true of gravity and both electrostatic and magnetic attractions. We may then write

$F_c = C/r^2$, where C is a constant. As F_c is equal and opposite to F_r , we have

$$\frac{C}{r^2} = 4\pi^2 n^2 m r.$$

$$\therefore n = \frac{1}{2\pi} \sqrt{\frac{C}{mr^3}}, \quad (1)$$

or

$$T = 2\pi \sqrt{\frac{mr^3}{C}}.$$

This shows us that in the case of the planets revolving around the sun, the period (length of a year) varies directly as the square root of the radius cubed. A shorter radius means a higher frequency, so that the planets nearest the sun have the shortest years.

89. Illustrations of central forces. The "banking" of a curved race track or railroad is a familiar illustration of the case where any centripetal force is possible with a given radius. When an automobile turns a curve, the wheels are held in the road by friction, but the inertia of the mass as a whole tends to keep it traveling in a straight line. The reaction to the force which holds the car in a circular path acts through its center of mass, and, combined with the force of friction F_c acting on the tread of the wheels, produces a couple tending to overturn it. This is shown in Fig. 64, where the car is supposed to be turning to the left, and the couple tends to overturn it to the right. That is, it tends to tip over *outside* the circle it is moving in. This effect can be completely neutralized by banking the road so that its outer edge is higher than the inner.

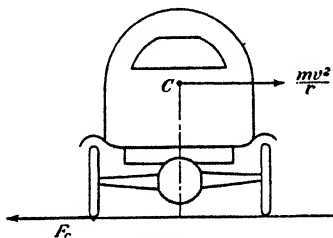


Fig. 64.

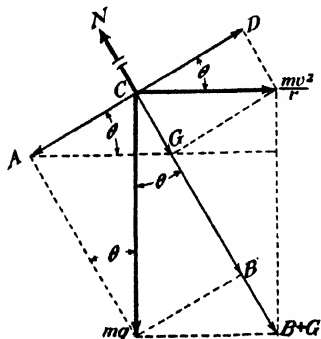


Fig. 65.

In Fig. 65, the force of gravity mg acts through the center of gravity C of the car. The road is banked so that the car is tilted at an angle θ . Then we may resolve mg into two components acting parallel and normal to the tilted profile of the road. These components are A and B re-

spectively. The centrifugal reaction mv^2/r may likewise be resolved into the components D and G , also parallel and perpendicular to the road's surface. The component D tends to overturn the car outward, while A tends to overturn it inward. If these are equal, there is no overturning tendency and the car bears equally on both right and left wheels. Thus we have made the car's weight take the place of F_c (Fig. 64) in holding it to the curved path. There is now no tendency to skid and as both D and A act through the center of gravity there is no overturning couple.

The equation by which we may find the proper banking angle for any assumed velocity and radius of curvature of the road is obtained by equating A and D , whence

$$mg \sin \theta = \frac{mv^2}{r} \cos \theta.$$

$$\therefore \theta = \tan^{-1} \frac{v^2}{gr}.$$

The normal force the car exerts on the road is $B + G$, which is equal to $mg/\cos \theta$, and is therefore greater than the weight of the car, as is evident from the diagram. The normal reaction N of the road, shown by the broken arrow, is equal and opposite to $B + G$. This vector and the two others represented by thick arrows would form a closed triangle, so they maintain the car in dynamic equilibrium.

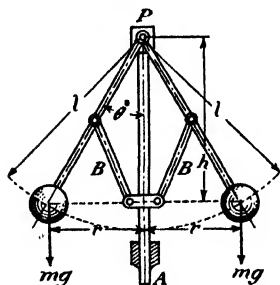


Fig. 66.

The flyball governor, or conical pendulum, is a familiar mechanism used to "govern" the speed of steam engines, and is shown in Fig. 66. The balls rotate on radii r, r , around the axis A and are so pivoted at P that they can move in an arc of a circle of radius l , thus varying the radius of the circle in which they are revolving.

The condition of equilibrium is the same as in the preceding illustration, although it is here more convenient to express the centrifugal reaction as $m\omega^2 r$. Therefore, as seen in Fig. 67, when $A = D$, $mg \sin \theta = m\omega^2 r \cos \theta$, or $\tan \theta = \omega^2 r/g$. But $\tan \theta = r/h$, so $r/h = \omega^2 r/g$.

$$\therefore n = \frac{1}{2\pi} \sqrt{\frac{g}{h}}, \text{ and } T = 2\pi \sqrt{\frac{h}{g}}.$$

This expression for the period will be seen later to be identical with the equation of the simple pendulum whose length is h . The governor

operates as follows: When the engine speeds up and T becomes too short, h decreases as the balls rise to rotate on a larger radius. This motion, acting through the links BB , is communicated to the valve which controls steam admission, and by partly closing it, causes the engine to slow down. When T is too long, h lengthens, and the reverse process takes place.

Let a liquid be rotated in a cylindrical vessel about its own vertical axis. The free surface of the liquid is no longer horizontal but becomes automatically "banked" as in the case of the curved road, thus balancing the central forces, and preventing radial motion. In Fig. 68 let P be a particle of mass m at the surface of the liquid which is revolving at that point about a radius r with an angular velocity ω . If the tangent to the surface at that point is DA inclined at an angle θ with the horizontal, then the force of gravity mg has a component A tending to move P down the slope, while $m\omega^2 r$ has a component D in the opposite direction. When A and D are equal, no motion results, and the surface is in kinetic equilibrium. We may then equate their values as before, obtaining, from $D = A$,

$$m\omega^2 r \cos \theta = mg \sin \theta.$$

$$\therefore \tan \theta = \frac{\omega^2 r}{g}.$$

Therefore the tangent of θ varies directly as the radius for a given angular velocity. As this is the property of a parabola, the curve shown is the section of a paraboloid of revolution and the surface of the liquid maintains the shape of a perfect parabolic reflector.

In the case of the centrifugal separator as used to take the cream out of milk, the situation is similar, but here the particles are of different densities and a selective process takes place. The heavier milk particles call for a greater centripetal force to hold them in a

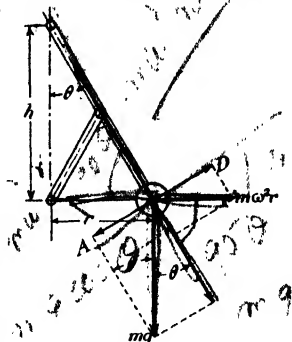


Fig. 67.

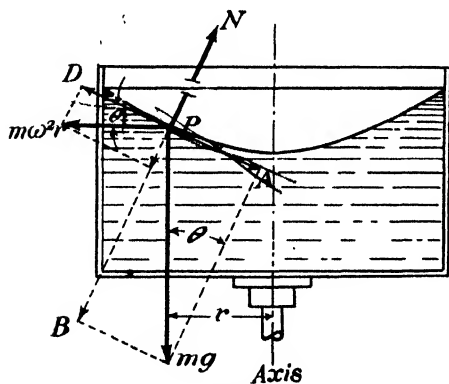


Fig. 68.

given orbit than does the lighter cream, and as this is not available they tend to concentrate near the circumference of the vessel, forcing the cream to collect around its axis, whence it is readily drawn off while the vessel is spinning.

Simple Harmonic Motion

90. Definition. When a particle revolves in a circular path with constant speed, its projection on any diameter moves back and forth along that diameter with a motion which is called **simple harmonic**. This is illustrated by the up-and-down motion of the crankpin of an engine's flywheel when seen edgewise, or the motion of the shadow it would cast on the floor if the sun were shining directly overhead. The vertical motion of a weight hanging at the end of an ideal helical spring and set oscillating up and down is also simple harmonic.

This motion is most conveniently studied through an analysis of the circular motion of a particle whose projection on a straight line performs simple harmonic vibrations. The particle and its circular path may then be called the **particle of reference** and the **circle of reference**, while its projection will be called the **vibrating particle**.

91. "Displacement" in simple harmonic motion. There are three important quantities involved in simple harmonic motion, to be expressed in terms of the time elapsed

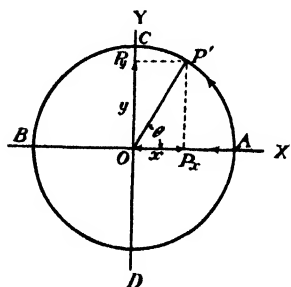


Fig. 69.

after the motion is supposed to have started. These are the displacement, the velocity, and the acceleration of the vibrating particle. By **displacement** is meant the distance of the vibrating particle P_x (Fig. 69) from its mean position which is taken as the origin. Suppose this particle is performing simple harmonic motion between A and B about O as the mean position. Then a circle

drawn with AB as a diameter and radius r is the circle of reference, and the particle P' revolving counterclockwise at constant speed is the particle of reference, with P_x always directly below it. The angle θ is the angle described by the radius vector OP' during the time t elapsed since P' last passed through A. But the angular velocity multiplied by the elapsed time gives the angle in radians, or $\theta = \omega t$. Therefore the displacement of P_x from the origin is given by

$$x = r \cos \theta = r \cos \omega t. \quad (1)$$

If P had been traveling along the Y diameter instead, then assuming time to have started with P' at A as before, the displacement y is given by

$$y = r \cos\left(\frac{\pi}{2} - \omega t\right),$$

where $\pi/2$ is 90° expressed in radians. Therefore

$$y = r \sin \theta = r \sin \omega t. \quad (2)$$

Equations (1) and (2) describe the displacement in simple harmonic motion for two important cases. In the first the vibrating particle began its motion at A , the extreme end of its path. Then $\theta = 0$, $\cos \theta = 1$, and x is a maximum. This largest value, r , of the displacement is called the **amplitude**. The angle θ is known as the **time angle** and is the same in both the cases we are considering. The time of a complete vibration from A back to A again is called the **period**, and is denoted by T . The number of complete vibrations per second is the **frequency**, denoted by n , the reciprocal of T .

In the second case, when $y = r \sin \omega t$, the particle starts at the origin and moving along the Y axis reaches C , its maximum displacement, when $\theta = \pi/2$, $\sin \pi/2 = 1$, and $y = r$. At the end of a period, P_y will be back at the origin again after an excursion to the extreme negative displacement at D , and will be about to start on its second complete vibration.

92. Graphic representation. The two harmonic motions along axes perpendicular to each other, but with the same particle of reference and the same angle, may be shown graphically by laying off the successive values of the X or Y displacements as ordinates against the

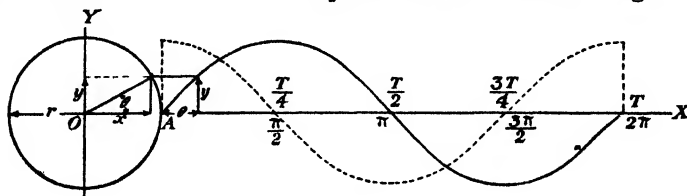


Fig. 70.

time angles, or fractions of the period, as abscissae. The solid line in Fig. 70 is the graph of $y = r \sin \theta$, and represents the instantaneous displacements of a particle moving harmonically on the Y axis. The method of construction is indicated. The magnitude of θ , measured by any convenient scale, is laid off on the X axis from A as an origin. The ordinates of the curve are equal to the Y displacements, and are given by $y = r \sin \theta$.

The dotted curve represents the first case, $x = r \cos \theta$, and is sim-

ilarly constructed, though the X displacements of P are here turned through 90° and laid off as ordinates for more convenient representation. The form of the curves is known as **sinusoidal**, and the curves are called **sinusoids**. They are much used in physics and engineering.

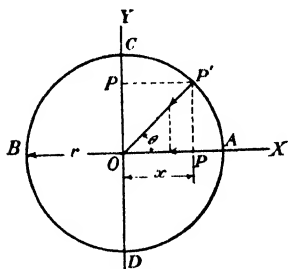


Fig. 71.

93. Acceleration and velocity in simple harmonic motion. In addition to displacement, it is important to determine both the velocity and acceleration of P at any time or position. Since P' in Fig. 71 is supposed to be moving with uniform angular velocity, its acceleration along the radius is given, as usual, by $a = \omega^2 r$. But the acceleration of P can be equal only to the projection of this value on the axes; therefore

$$a_x = -\omega^2 r \cos \theta = -\omega^2 r \cos \omega t,$$

and

$$a_y = -\omega^2 r \sin \theta = -\omega^2 r \sin \omega t,$$

which is the acceleration of P in terms of the time elapsed since the particle was at A or O , and the negative sign indicates that it is directed toward the origin, or opposite to the displacement. Also, since $\cos \theta = x/r$, and $\sin \theta = y/r$, both a_x and a_y may be found in terms of the displacement, or

$$a_x = -\omega^2 r x / r = -\omega^2 x, \quad (1)$$

and

$$a_y = -\omega^2 r y / r = -\omega^2 y,$$

which are very useful expressions.

As ω is constant, they show that in simple harmonic motion the acceleration is always directed toward the middle of the path, and is proportional to the displacement. This is the essential criterion of this kind of motion, that is, the motion is always s.h.m. (simple harmonic motion) whenever a varies as x or y , and is directed toward the origin.

The velocity of P at any time angle θ is obtained by resolving the linear velocity of P' into vertical and horizontal components as in Fig. 72. The former is the Y velocity of P , the latter its X value,

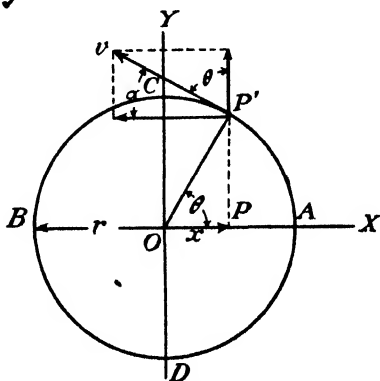


Fig. 72.

and they are equal to $v \cos \theta$ and $v \sin \theta$ respectively. But linear velocity in a circle is equal to ωr ; therefore

$$v_x = -\omega r \sin \theta = -\omega r \sin \omega t, \quad (2)$$

and

$$v_y = +\omega r \cos \theta = +\omega r \cos \omega t,$$

from which v_x and v_y may be found at a given time.

From a consideration of equations (2), as well as from actual observation, the velocity is seen to be a maximum when the vibrating particle is passing through the origin. It is then that $\sin \omega t = 1$ for horizontal, and $\cos \omega t = 1$ for vertical vibrations, giving ωr as the maximum value in both cases. When P is at the end of its swing, the velocity is 0, for then $\theta = 0$ in the former case, and $\theta = 90^\circ$ in the latter.

These extreme values, it should be noticed, are contrary to those of the acceleration, for a is zero at the origin, and is maximum at maximum displacement. This seems paradoxical at first sight, for it might seem as if, when a particle stops as it does at the end of its swing, its acceleration must be zero. But the reversal of direction at A or C involves a change in the velocity, even if its instantaneous value passes through zero during that change.

94. Period of simple harmonic motion. From the equation giving the acceleration at any point distant x or y from the origin, it is possible to calculate the period of vibration in terms of the displacement and acceleration. For

$$a = -\omega^2 x = -\omega^2 y. \quad (1)$$

$$a = -\frac{4\pi^2 x}{T^2} = -\frac{4\pi^2 y}{T^2}, \quad (2)$$

and

$$T = 2\pi\sqrt{\frac{-x}{a}} = 2\pi\sqrt{\frac{-y}{a}}. \quad (3)$$

The negative sign under the radical does not yield an imaginary value of the period, because when x or y have positive values leaving the numerator negative, the acceleration is negative. And when x and y are negative, the numerator becomes positive and a is positive also.

95. The simple pendulum. A nearly perfect illustration of simple harmonic motion is obtained by swinging a small dense mass, or "bob," at the end of a light string. Such an arrangement when the dimensions of the bob and mass of the string are negligible is known as a **simple pendulum**. If the arc through which it vibrates is small,

the motion is a very close approximation to harmonic, along a circular instead of a straight path as in the preceding cases. When the pendulum whose length is l is in the position shown in Fig. 73, the force of gravity may be resolved into two components. One of these,

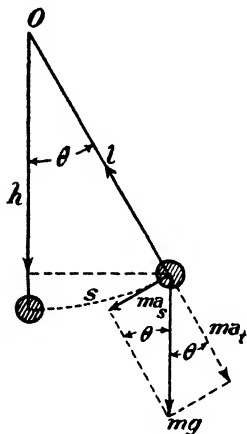


Fig. 73.

ma_s , is equal to $mg \cos \theta$ acting along the string and produces tension, while the other, the tangential component ma_t , tends to restore the pendulum to its position of rest, and is equal to $-mg \sin \theta$. The negative sign means that the force tends to diminish the distance from the point of rest. Then, taking $ma_s = -mg \sin \theta$, and dividing by m , we obtain $a_s = -g \sin \theta$. If θ is very small, say less than 8° , its value, s/l , in radians is approximately equal to its sine, or $\sin \theta = s/l$, whence $g \sin \theta = gs/l = -a_s$. The acceleration is therefore proportional to the displacement s , is directed toward the position of rest, and the motion is simple harmonic. But it has just been proved (equation 1, Article 93) that in this case

$a = -\omega^2 x$, where x is the displacement. Therefore, substituting s for x , and $4\pi^2/T^2$ for ω^2 , we obtain

$$\frac{gs}{l} = \frac{4\pi^2 s}{T^2},$$

and

$$T = 2\pi\sqrt{\frac{l}{g}}.$$

Thus it appears that for small angles, the period of such a pendulum is independent both of the amplitude of vibration and of the mass of the bob. The period of the conical pendulum or flyball governor, already described, was found to be given by an equation similar to the above, but with h in place of l . In this case no approximation was necessary, and the value of T was rigorous for all angles.

96. Energy of a swinging pendulum. Before leaving the simple pendulum, it is well to note the remarkable illustration of the conservation of energy which it affords. At the beginning of the swing, with the string vertical, the bob has a horizontal velocity, which as we have seen is a maximum at that point. The potential energy of the system with respect to the position of rest is zero. As the bob moves outward, it loses velocity and therefore kinetic energy also, but gains potential energy as it rises above its original level. At the

extreme end of its swing, the energy is all potential, and equal to $mg(l - h)$. This may be equated to the original kinetic energy $\frac{1}{2}mv^2$, where v is the maximum velocity when the pendulum passes through the position of rest. At intermediate positions the energy is partly of one kind and partly of the other, but their sum is always constant, as may be proved by calculation, provided there is no loss and consequent slowing down due to friction and the viscosity of the air.

SUPPLEMENTARY READING

H. A. Erickson, *Elements of Mechanics* (Chap. 5), McGraw-Hill, 1927.

J. B. Reynolds, *Elementary Mechanics* (Chap. 7), Prentice-Hall, 1928.

PROBLEMS

1. Calculate the angular velocity in radians per sec. of a particle which makes 300 revolutions per min. What is the linear velocity if the radius is 4 ft.? *Ans.* 31.4 ra./sec.; 125.6 ft./sec.

2. What is the uniform angular acceleration of a particle moving in a circle if, starting from rest, it reaches 240 r.p.m. in 30 sec.? What is its linear acceleration if the radius is 2 ft.? *Ans.* 0.84 ra./sec²; 1.68 ft./sec².

3. Calculate the centripetal acceleration of the particle in Problem 1. *Ans.* 3944 ft./sec².

4. The centripetal acceleration of a particle moving in a circle of 80 cm radius is 400 cm/sec². What is its period? *Ans.* 2.8 sec.

5. A stone weighing 200 g is whirled around in a horizontal circle at the end of a string 60 cm long with an angular velocity of 150 r.p.m. Calculate the tension in the string. *Ans.* 2.96 megadynes.

6. A string which can just support a tension of 10 lb. breaks when a stone weighing 8 oz. is attached to it and whirled around in a horizontal circle whose radius is 18 in. What was the period of revolution? *Ans.* 0.30 sec.

7. What would be the length of the year if the earth were half its present distance from the sun? *Ans.* 129 days.

8. What is the proper angle for banking a road around a curve of 200 ft. radius to allow for speeds of 40 miles per hour? *Ans.* 28°1.

9. What is the proper speed for an automobile rounding a curve of 135.3 ft. radius and banked at 30° with the horizontal? (Take $g = 32$ ft./sec².) *Ans.* 50 ft./sec.

10. A particle starting at the end of its swing performs simple harmonic vibrations. It has an amplitude of 12 cm and a frequency of 40 vibrations per minute. What is the displacement at the end of 2 sec.? *Ans.* -6 cm.

11. If the particle in Problem 10 starts from the middle of its swing, what will be the displacement at the end of 2 sec.? *Ans.* +10.38 cm.

12. What is the acceleration of the particle in Problems 10 and 11?
Ans. 105.2 cm/sec².; -182 cm/sec².

13. Calculate the velocities of the particle in Problems 10 and 11. What is the maximum velocity? *Ans.* -43.5 cm/sec.; -25.1 cm/sec.; 50.2 cm/sec.

14. A vibrating particle has an acceleration of 9 cm/sec². when its displacement is 12 cm. Calculate the period. *Ans.* 7.26 sec.

15. A simple pendulum whose length is 36 cm makes 49.8 complete vibrations per minute. Calculate g . *Ans.* 978 cm/sec².

16. What is the length of a simple "seconds pendulum" whose half period is 1 sec., where $g = 980$ cm/sec²? *Ans.* 99.3 cm.

***17.** If a clock intended to be regulated by a seconds pendulum has a pendulum only 90 cm long, how much will it gain in a day? *Ans.* 1 hr., 12 min., 30 sec.

18. A stone weighing 10 lb. is whirled in a vertical circle of 2 ft. radius. What is the least possible angular velocity in r.p.m.? What is the pull on the string when the stone is at the bottom of its path? (use $g = 32$ ft./sec²).
Ans. 38.22 r.p.m.; 20 lb.

CHAPTER 7

Rotation of a Body

97. Moment of inertia. Suppose you had a set of wheels of different weights and sizes, some like solid discs, and others with light spokes like bicycle wheels, but all mounted on ball bearings so as to rotate with little friction. Then suppose you experimented with them in an attempt to find out how Newton's laws apply to rotation. You would at once observe that rotating bodies have inertia and tend to keep on spinning after having been started, and that a large torque accelerates or stops them faster than a small one. You would also notice that heavy wheels have more of this inertia than light ones, and that a wheel with most of its mass in the rim is harder to start than one with the mass evenly distributed. Finally, if you were very inquisitive and held the axis at its two ends, you would notice that when spinning, the wheel resists any tilt of its axis in a most extraordinary way. This is equivalent to changing the direction of a mass moving in a straight line and is called the gyroscopic effect which we shall consider further on.

The inertia of a massive wheel is illustrated by the flywheel of a steam engine which carries the engine over its "dead center" and smooths out irregularities of effort on the crankpin. If the effort increases, this rotational inertia tends to prevent speeding up, and if it decreases, the inertia prevents slowing down.

As rotations are produced by the moment of a force, the kinetic reaction against this force must be in the nature of a moment also. Therefore the property of a body to which this reaction is due is known as its **moment of inertia**, usually represented by the letter I .

We noted in our supposed experiments that the moment of inertia is greatest when the mass is farthest from the axis. This is because when the radius is greater the mass has a higher linear velocity with the same angular velocity. From this we conclude that the moment of inertia must involve both the rotating mass and its distance from the axis. The exact way in which these quantities are related is found as follows: Referring to Fig. 74, suppose several different particles of masses $m_1, m_2, \dots m_n$ rotate about a common axis O ,

on radii r_1, r_2, \dots, r_n and with a common angular velocity. Then their linear velocities at any instant are $\omega r_1, \omega r_2, \omega r_3, \dots, \omega r_n$, and their kinetic energies are $\frac{1}{2}m_1\omega^2r_1^2, \frac{1}{2}m_2\omega^2r_2^2, \dots, \frac{1}{2}m_n\omega^2r_n^2$, giving as the total energy of the system

$$\begin{aligned} W &= \frac{1}{2}\omega^2(m_1r_1^2 + m_2r_2^2 + m_3r_3^2 + \dots + m_nr_n^2) \\ &= \frac{1}{2}\omega^2\Sigma mr^2. \end{aligned} \quad (1)$$

If this is compared with the usual $\frac{1}{2}v^2m$ for linear kinetic energy it is obvious that v^2 has been replaced by ω^2 , and m by Σmr^2 which is thus seen to be the rotational analogue of mass, and is therefore the moment of inertia I of the system. It is for rotation what m is for translation, and its dimensions are $[ML^2]$.

The kinetic energy of a rotating body is then given in the absolute system of units by

$$W = \frac{1}{2}I\omega^2. \quad (2)$$

But in the gravitational system where the constant k is not unity, the linear kinetic energy is $\frac{1}{2}mv^2/k$ as we have seen; therefore equation (1) becomes

$$W_g = \frac{1}{2} \frac{\omega^2(\Sigma mr^2)}{k} \quad (3)$$

$$= \frac{1}{2} \frac{I\omega^2}{k}. \quad (4)$$

Equation (1) gives kinetic energy in ergs if m and r are expressed in grams and centimeters, and in foot-pounds if they are expressed in pounds and feet. Equation (3) gives kinetic energy in gram-centimeters or foot-pounds with the same choice of units, and with k equal approximately to 980 in one case and 32.2 in the other.

In computing I , if masses are measured in slugs, this amounts to dividing by $k = 32.2$, and the resulting value I_g gives the kinetic energy in gravitational units, using equation (2) with I_g substituted for I .

98. Calculation of moments of inertia. In the ideal system just used it is extremely easy to calculate the numerical value of I provided each mass and its distance from the axis are given; but when a real body made up of innumerable mass particles rotates about an axis, it is necessary to consider all these infinitesimal masses, each revolving in its own special orbit. Generally the problem is insoluble, and I can be determined only by experiment. But in a few cases of

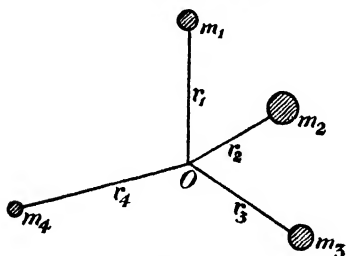


Fig. 74.

a homogeneous body of some simple regular shape, the moment of inertia with reference to a limited choice of axes may be found by the methods of the calculus.

In the following special forms, I has been obtained in this manner and its values are here given for reference:

A solid cylinder of radius r about its own axis.



$$I = \frac{mr^2}{2}$$

A cylindrical ring of rectangular section, whose outer radius is r_1 and the inner radius is r_2 , about its own axis.



$$I = m \left(\frac{r_1^2 + r_2^2}{2} \right)$$

A solid cylinder of length l about an axis normal to its own, through the center.



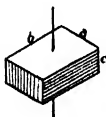
$$I = m \left(\frac{r^2}{4} + \frac{l^2}{12} \right)$$

A sphere about any diameter.



$$I = \frac{2}{5} mr^2$$

A rectangular parallelopiped of edges a , b , c , about an axis normal to the ab face, at its center.



$$I = m \left(\frac{a^2 + b^2}{12} \right)$$

The second and last of these have ideal cases, which are sometimes useful. The cylindrical ring may have such a small thickness that the two radii may be regarded as equal, an approximation nearly realized in the flat iron tire of a cart wheel. Then $r_1^2 + r_2^2 = 2r^2$ and $I = mr^2$, which indeed follows directly from the definition $I = \Sigma mr^2$, because all the masses are practically equidistant from the center.

Also, if the parallelopiped is in the form of a long slender strip, whose depth c may have any value, but with its width a small compared to its length b , then the moment of inertia reduces to

$$I = \frac{mb^2}{12}, \text{ or } \frac{ml^2}{12}$$

as it is usually written.

99. Shifting the axis. In all the preceding formulae the axis of rotation passes through the center of mass of the body, but it is possible to obtain I about any other axis parallel to it at a distance h by the following equation, known as Steiner's theorem:

$$I = I_c + mh^2,$$

where I_c is the moment of inertia about an axis through the center of mass, h is the perpendicular distance between the axes, I is the moment of inertia required and m is the total mass.

As an illustration of Steiner's theorem, suppose we wish to find the moment of inertia of a grindstone mounted eccentrically on a shaft 20 cm from its center, when its mass is 20 kg and its radius 40 cm. Then $I_c = \frac{1}{2}mr^2 = 10,000 \times 1600 = 16 \times 10^6 \text{g-cm}^2$ and $I = 16 \times 10^6 + 20,000 \times 20^2 = 24 \times 10^6 \text{g-cm}^2$.

100. Rotational actions and reactions. The rotational analogue of a force, as has been stated, is the moment of that force. The moment varies with its lever arm as well as with the force itself, and is therefore measured by their product Fr .

In the case of pure translation the reaction to an action is either an equal and opposite force of the same kind, or it is the kinetic reaction measured by ma when change of motion is taking place. In either case $\Sigma A = 0$.

In the case of rotation there is a similar relation. But now moments or torques due to couples take the place of forces and we must write instead $\Sigma L = 0$. If there is no motion or only uniform motion when a body is subjected to a moment or torque L , there must be an equal and opposite reactive moment or torque L' which opposes L ; therefore $L - L' = 0$. If change of motion results, then there is a rotational kinetic reaction exactly analogous to the linear kinetic reaction of pure translation. This is calculated by taking the product of the moment of inertia and angular acceleration, or $I\alpha$, instead of the product of mass and linear acceleration, ma , so that we have $L - I\alpha = 0$ instead of $F - ma = 0$.

The fact that the reactive torque equals $I\alpha$ may be roughly tested by trying to turn over a heavy flywheel and get it going fast. A push of say one hundred pounds applied tangentially to its rim is about the best a man can do, and if its rim is heavy, he finds it harder to speed it up than if its mass were small and evenly distributed, as in a wooden disc. The heavy rim means that I is large and the reactive torque is large also. Similarly it takes a greater force to speed up the wheel rapidly than it would to do so slowly. A massive flywheel rotating without friction might be got going at sixty revolutions per minute by a weak but long continued torque, but it takes a great torque to speed it up rapidly.

This very important principle that $L' = I\alpha$ may be proved rigorously as follows: The linear kinetic reaction of a mass m at a distance of r cm from the axis is ma , and the moment of this reaction

is $mra = L'$. But the linear acceleration equals αr ; therefore, substituting αr for a , we obtain

$$L' = mr^2\alpha,$$

or in the gravitational system,

$$L'_g = \frac{mr^2\alpha}{k},$$

which is the reaction against the torque causing the angular acceleration α . Then we may sum up all the reactions of all the infinitesimal masses of the rotating body, remembering that α is the same for all, so that

$$L' = \alpha \Sigma mr^2 = \alpha I,$$

and in the gravitational system,

$$L'_g = \alpha I_g = \frac{\alpha I}{k}.$$

101. Moment of momentum. The momentum of a rotating body, like torque and the kinetic energy of rotation, is angular rather than linear, and may be obtained from the linear momentum of an individual particle of mass m , moving in a circle. This is mv . The moment of this quantity is mvr , and since $v = \omega r$, $mvr = \omega mr^2$. But ω is constant for all the particles of which a rigid body is composed. Therefore, summing them up as before, we find that the moment of momentum or angular momentum of the body is

$$\omega \Sigma mr^2 = I\omega,$$

and in the gravitational system,

$$\omega \Sigma \frac{mr^2}{k} = I_g \omega = \frac{I\omega}{k}.$$

The dimensions of this quantity are found from those of

$$[\omega] = [T^{-1}], \text{ and } [I] = [ML^2].$$

Hence

$$[I\omega] = [ML^2T^{-1}].$$

A comparison with the dimensions of energy shows that the above are those of energy multiplied by time, and the unit of angular momentum is therefore an erg-second. It is much used in the quantum theory of the energy of radiation.

Like linear momentum, angular momentum remains constant unless some external action (torque in this case) changes it. Thus a revolving system having a configuration which can be altered internally so as to vary I , experiences an increased angular velocity if I

decreases, or a decrease in ω if I increases, but the product $I\omega$ remains constant.

A well-known experiment strikingly illustrates the conservation of angular momentum as follows: Let the student stand on a small platform that can be made to rotate about a vertical axis, and let him hold out two fairly heavy weights at arm's length while the platform is set slowly spinning. Then if he lowers his arms, his moment of inertia decreases, so that the angular velocity must increase in order to keep $I\omega$ constant. The result is very convincing, as ω increases far too much for comfort, so that the experimenter hastens to raise his arms again in order to slow down to a less giddy rate of rotation.

102. Comparison of rotation with translation. Before proceeding to certain practical applications of the theory of rotation, it will be well to compare the quantities it makes use of with the corresponding quantities in translation. The following table gives these analogous quantities in parallel columns, with a third column in which will be found their relation to each other.

Translation	Rotation	Relation
distance..... s	angle..... θ	$\theta' = \text{arc/radius} = s/r$
time..... t	time..... t	$t = t$
mass..... m	moment of inertia..... I	$I = \Sigma mr^2$
velocity..... v	angular velocity..... ω	$\omega = v/r$
acceleration..... a	angular acceleration..... α	$\alpha = a/r$
force..... F	moment, torque..... L	$L = Fr$
work..... $W = Fs$	work..... $W = L\theta$	$L\theta = Fs$
kinetic energy $W = \frac{1}{2}mv^2$	kinetic energy $W = \frac{1}{2}I\omega^2$	For mass-particles: $\frac{1}{2}I\omega^2 = \frac{1}{2}mv^2$ †
momentum..... mv	moment of momentum $I\omega$	$I\omega = mv r$

Equations relating to translation become true for rotation provided the corresponding quantities are substituted. Thus in accelerated motion:

$$v = at \quad \text{becomes} \quad \omega = \alpha t$$

$$s = \frac{1}{2}at^2 \quad \text{"} \quad \theta = \frac{1}{2}\alpha t^2$$

$$v^2 = 2as \quad \text{"} \quad \omega^2 = 2\alpha\theta$$

$$F = ma \quad \text{"} \quad L = I\alpha$$

It should however be noted that only time and work or energy, are the same sort of quantity in both kinds of motion, and so can be

† Also nearly true when the mass of a wheel is concentrated mainly in the rim.

added together or equated. The others are only *corresponding* quantities, and cannot be added or equated.

103. Radius of gyration. If we were to concentrate the entire mass of a rotating body at a single point, the kinetic energy of rotation for the same angular velocity would in general be altered. But it is possible to find a point, or series of points in a circle around the axis, where the mass could be concentrated without altering the kinetic energy of rotation. Thus a thin hoop of radius K , or a particle at a distance K from the axis could be made to have the same kinetic energy as a solid disc of the same weight rotating at the same speed. The kinetic energy of such a hoop is given by $\frac{1}{2}MK^2\omega^2$, and setting this equal to the kinetic energy of the disc or other body rotating about the same axis, we have

$$\frac{1}{2}MK^2\omega^2 = \frac{1}{2}I\omega^2,$$

whence

$$K = \sqrt{\frac{I}{M}}.$$

This quantity K is known as the **radius of gyration**, and when it is known, we may use the product MK^2 , instead of I in all problems concerning rotation.

104. Problems involving rotation. Suppose a solid disc, like a grindstone, rotates about a horizontal axis, and it is required to find how long a tangential force applied to its rim by a brake will take to stop it. The equation for the corresponding case in translation is $Ft = mv$. Therefore we may write for rotation, $Lt = I\omega$, and if the radius, moment of inertia, and revolutions per second are known, the problem is readily solved. If, however, the angle (or number of revolutions) before it stops is called for, the equation corresponding to $Fs = \frac{1}{2}mv^2$ is $L\theta = \frac{1}{2}I\omega^2$, or $L_g\theta = \frac{1}{2}I\omega^2/k$ in the gravitational system where L_g is measured in foot-pounds or gram-centimeters. In either case, θ is measured in radians, but this result divided by 2π gives the number of revolutions required.

When a disc rolls along a plane, the problem involves both translation and rotation. The kinetic energy is equal to $\frac{1}{2}mv^2 + \frac{1}{2}I\omega^2$ and if a steady force F acting horizontally through its center is applied to stop it, the distance through which it must act is given by

$$Fs = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2. \quad (1)$$

In this equation we may find the angular velocity from the linear velocity of the center by the usual relation $\omega = v/r$. This is because the wheel rolls a distance $2\pi r$ in the time T of one revolution,

$v = 2\pi r/T$, and since it rotates through 2π radians in one revolution, $\omega = 2\pi/T$, and $\omega = v/r$. This is the same relation as that obtained when a wheel rotates about a fixed axis and v is the velocity of a point on its rim. Therefore in the case of rolling, the velocity of the center takes the place of the peripheral velocity of a pure rotation.

As an illustration of these principles applied to a rolling disc, let $m = 10$ kg, and $v = 2$ m per sec. Required, the distance the disc rolls while a force of a million dynes acting horizontally through its center opposes it.

Since I for this solid disc is $mr^2/2$, and $\omega = v/r$, we obtain from (1)

$$Fs = \frac{1}{2}mv^2 + \frac{1}{4}mv^2 = \frac{3}{4}mv^2,$$

and

$$s = \frac{3mv^2}{4F} = \frac{3 \times 10^4 \times 4 \times 10^4}{4 \times 10^6} = 300 \text{ cm.}$$

This calculation shows the value of solving a problem symbolically as far as possible before introducing numerical values. If this had not been done the labor would have been much greater, and it would have been necessary to know r in order to find I .

When a solid disc rolls down an inclined plane, as shown in Fig. 75, the original potential energy mgh at the top is converted into kinetic energy both of translation and rotation. At the bottom the total kinetic energy gained equals the potential energy lost, or

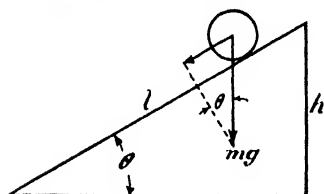


Fig. 75.

$$mgh = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2. \quad (2)$$

Since part of the potential energy goes into the kinetic energy of rotation, less of it is available for the energy of translation. Therefore the disc gathers speed more slowly than if it could slide without friction, and the greater its moment of inertia with a given mass, the smaller will be its velocity of translation v . Thus a hoop rolls more slowly down an incline than a solid disc of the same mass and radius, because its moment of inertia is twice as large.

The time of descent of a disc rolling down an incline, the acquired velocity, and the distance covered in a given time can be found as follows: In Fig. 75, it is required to find the velocity of the disc after rolling from the top to the bottom of the plane. In the preceding problem it was proved that the total kinetic energy of such a disc is given by $\frac{3}{4}mv^2$. This must be equal to the potential energy it had at the top of the slope, which was mgh . Therefore, equating these two

kinds of energy, we find $v = 2\sqrt{gh/3}$, which is an expression independent both of the mass and the radius of the disc.

As the body started from rest, the average velocity is one half this value, and the time required is l/v_{av} ; hence $t = l\sqrt{3/gh}$. But for *any* distance s , the corresponding height is $s \sin \theta$; hence

$$t = s\sqrt{\frac{3}{gs \sin \theta}},$$

and

$$s = \frac{1}{3}gt^2 \sin \theta,$$

which is the distance *rolled* in t seconds. This should be compared with $s = \frac{1}{2}gt^2 \sin \theta$, for freely *sliding* bodies.

105. Harmonic motion of bodies (translation). If the motion is linear without rotation, the body may be treated as a particle. Thus a mass m hung at the end of an ideal helical spring performs simple harmonic vibrations after being pulled downward and then released. In the case of such a spring the force of restitution is directly proportional to the displacement y from rest, or $F = -cy$ where c is the constant of proportionality. That is, double pull means double stretch, triple pull means triple stretch, and so on. Then if pulled down from its position of rest through a distance y and released, the kinetic reaction ma is equal to the above force, so that $ma = -cy$ or $-y/a = m/c$. Therefore, since m and c are constants, a varies as y , is oppositely directed, and the motion is harmonic. We may then use the general equation for the period of any harmonic vibration (equation 3, Article 94), which is,

$$T = 2\pi\sqrt{\frac{-y}{a}}.$$

But

$$-\frac{y}{a} = \frac{m}{c}.$$

$$\therefore T = 2\pi\sqrt{\frac{m}{c}}, \quad (1)$$

where c , the elastic constant of the spring, may be called the coefficient of restitution. It is that quantity by which the displacement must be multiplied to obtain the restoring force.

When the mass m is at rest, the displacement l is due to the pull of gravity alone; therefore $cl = mg$. Thus a second value for T may be found by setting $m/c = l/g$, or

$$T = 2\pi\sqrt{\frac{l}{g}}. \quad (2)$$

Equation (1) enables us to calculate the constant of the spring in terms of the period of vibration, and from equation (2) the acceleration due to gravity may be found.

106. Harmonic motion of bodies (rotation). A body suspended by a wire, when given an initial twist about it as an axis, executes rotary vibrations which are simple harmonic. As the wire twists and untwists it obeys a law similar to $F = -cy$ of the helical spring. But now we are dealing with *torque* and *angular* displacement, so that the torsional equivalent is $L = -c\theta$, where c is the torsional coefficient of restitution of the wire, and is constant for moderate angles.

The rotational kinetic reaction to the torque is $I\alpha$, instead of ma . Therefore $I\alpha = -c\theta$, and since the moment of inertia of the body about a given axis, as well as c , is a constant, the angular acceleration is seen to be proportional to the angle and oppositely directed. Therefore the motion is simple harmonic, and we may apply the equation for the period of such motion, in which θ replaces s , and α replaces a . Then

$$T = 2\pi\sqrt{\frac{-\theta}{\alpha}}, \text{ and since } -\frac{\theta}{\alpha} = \frac{I}{c}, \quad (1)$$

$$T = 2\pi\sqrt{\frac{I}{c}}. \quad (2)$$

Here c applies to rotation instead of to translation, and I replaces m in the corresponding formula for linear vibrations. We may therefore calculate the period of any harmonic vibration from one of these two analogous formulae, provided we can determine the coefficient c , which when multiplied by the displacement (distance or angle) gives the corresponding reaction (force or torque). The generalized formula may then be expressed by

$$T = 2\pi\sqrt{\frac{J}{c}}, \quad (3)$$

where J means either mass or moment of inertia, according to circumstances, and c is either the linear or torsional coefficient of restitution.

107. The torsion pendulum. Any mass oscillating about a suspending wire as an axis maintains a constant period independent of θ , if the latter is not large enough to exceed the limits within

which c is constant. It is therefore known as a torsion pendulum. In the form shown in Fig. 76, it is used in some makes of clocks,

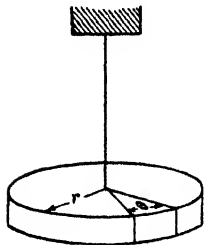


Fig. 76.

and its period can be calculated from the known moment of inertia of a solid disc about its own axis, provided c is known also. Then

$$T = 2\pi\sqrt{\frac{mr^2}{2c}} = \pi r\sqrt{\frac{2m}{c}}.$$

108. The simple pendulum. We may now calculate the period of this ideal arrangement from the point of view of rotation, using the general equation $T = 2\pi\sqrt{I/c}$. By referring to Fig. 77, it is seen that $I = ml^2$. The component of mg which acts as the restoring force is $mg \sin \theta$, and its moment L is $-mgl \sin \theta$. For small angles $\sin \theta = \theta$ nearly; therefore $L = -mgl\theta$. But mgl is constant, and we may write $L = -c\theta$, where c is the coefficient of restitution, because when multiplied by θ it gives the restoring moment. It is equal to mgl in this particular case provided θ is so small that we may regard the angle equal to its sine without sensible error.

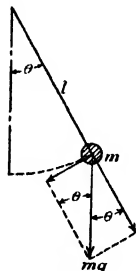


Fig. 77.

Then substituting ml^2 for I in the period equation, and mgl for c , we obtain, as before,

$$T = 2\pi\sqrt{\frac{ml^2}{mgl}} = 2\pi\sqrt{\frac{l}{g}}.$$

109. The compound pendulum. Any body vibrating in a vertical plane about a horizontal axis, with gravity supplying the restoring moment, is called a compound or physical pendulum. In Fig. 78, an irregular solid is shown deflected through an angle θ , against the restoring moment of gravity which acts through the center of mass at C , and causes the body to oscillate about its axis at O . The value of the moment is given by $L = -mgh\theta$ for small angles, where h is the distance between the center of mass and the axis measured in the plane of vibration, and the coefficient of restitution c is therefore mgh , so that

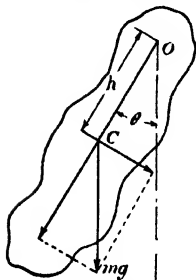


Fig. 78.

$$T = 2\pi\sqrt{\frac{I}{mgh}}.$$

110. Center of percussion. A simple pendulum of length l which has the same period as a compound pendulum is called an **equivalent simple pendulum**. If a compound pendulum is hanging at rest, and

a vertical line is drawn down through its axis, then a point on this line at a distance l from the axis is known as the **center of percussion**, or **center of oscillation** of the body. If a blow is struck at the center of percussion in a direction perpendicular to the axis, it produces pure rotation with no tendency to knock the pendulum off its support.

Such a point is the one at which the ball should come in contact with the bat in such games as baseball and cricket. If it does, there is no sense of shock on the batter's hands, as is the case when contact with the ball is made elsewhere. Since the distance of the center of percussion from the axis is equal to the length of a simple pendulum whose period is the same as that of the vibrating body, it is readily located by equating the periods of the simple and compound pendulums. Then

$$2\pi\sqrt{\frac{l}{g}} = 2\pi\sqrt{\frac{I}{mgh}}, \quad (1)$$

and
$$\frac{l}{g} = \frac{I}{mgh}.$$

Hence
$$l = \frac{I}{mh}, \quad (2)$$

where I is the moment of inertia of the body, m is its mass, and h is the distance between its axis and center of gravity.

There are then three important points to be found with regard to rigid bodies rotating about an axis. They are the center of gravity, the end of the radius of gyration, and the center of percussion. Two of these may coincide, but in general all three have different positions.

111. Vectors of rotations. The most significant direction which concerns rotation is that of the axis. No other direction has any real meaning, since every possible direction of motion in a plane normal to the axis is represented by the different particles of a rotating body. To represent, then, such quantities as ω , α , torque, and angular momentum as vectors, they are laid off along the axis of rotation. As usual, the vector's magnitude measures the numerical value of such quantities, and the *sense* of the vector is chosen so that it points away from the observer when the rotating body is viewed in such a direction that its sense of rotation appears clockwise. Thus the vector representing the rotation of a right-handed screw points in the direction in which it advances.

112. Rotational reactions. In Article 97, we proposed that the reader try holding two ends of the axis of a rapidly spinning wheel

and then try tilting it. He will find that it resists any attempt to change the direction of its axis, or, what is the same thing, of the plane in which it is rotating. This reaction to a torque applied to the axis is increased by increasing the angular momentum of the rotating body. That is, the reaction depends upon both angular velocity and moment of inertia, and is in the nature of a torque. This torque results in a shift of the axis in a plane normal to that in which we attempt to turn it, as is explained below. Vibrating bodies also resist a change of the plane in which they are vibrating in a manner similar to rotating bodies.

113. The gyroscope. A heavy mass of large moment of inertia set rotating at a high speed about an axis is called a **gyroscope**, and the group of phenomena associated with it is called **gyroscopic**.

Let the vector A in Fig. 79 represent the angular momentum $I_1\omega_1$ of the disc D of mass M , whose axle is pivoted at P with a universal ball and socket joint. Obviously D tends to fall under the force of gravity. This force combined with the reaction at P creates a torque equal to Mgl . But if D is spinning about its axis, this torque involves a change in its angular momentum due to a shift of the plane of rotation. If now the torque is applied for t seconds, a new angular momentum develops, given by $I_2\omega_2 = Lt$, as a result of rotation in the direction of the dotted arrow about an axis through P normal to and outward from the plane of the diagram. Therefore, viewed from above, this vector B appears perpendicular to the vector A , which is that of the angular momentum of the spinning disc. According to the right-handed screw convention, these vectors appear as in Fig. 80, where A is made longer than B to indicate a greater angular momentum of the rotating disc about its axis than that due to the torque exerted by gravity and the reaction of the support. The resultant

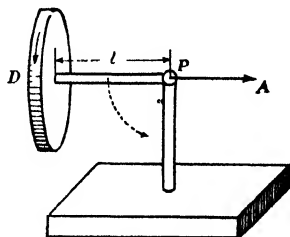


Fig. 79.

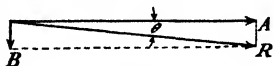


Fig. 80.

momentum, at the instant considered, lies along a new axis R , and the axle tends to shift into this new position, causing the gyroscope to rotate about P in a clockwise direction when seen from above.† The next moment, A (owing to the torque caused by gravity), is no

† This discussion is very incomplete, and applies only to the instantaneous position given in Fig. 79.

longer horizontal, so that its downward motion, combined with the horizontal rotation just explained, causes it to sweep over a kind of conical surface of steadily increasing pitch as the disc gradually descends.

This shifting of the axis is called **precession**. It is applied to the slow change in the direction of the earth's axis, called "precession of the equinoxes," which is due to the same causes as in the case cited above.

114. Foucault's pendulum. Angular momentum, like linear, resists change either in direction or magnitude, but the direction is precisely that of the vector representing it, along its axis. Therefore bodies, whether rotating or vibrating, resist a change of the direction of the axis, or, what is the same thing, the plane of rotation or vibration. When such a change is forced upon them, it involves a rotatory kinetic reaction.

A heavy mass of large inertia set swinging at the end of a long wire may be used to illustrate the tendency of bodies to continue vibrating in one plane. This fact was discovered by Foucault in 1851, and used by him to demonstrate the rotation of the earth on its axis. Such a pendulum at the North Pole would continue swinging in a fixed plane while the earth turned under it, so that to an observer the path of the swinging bob would seem to turn through 360° in 24 hours, or 15° per hour. At the equator the plane of vibration and that of the meridian rotate together and there is no relative change. At intermediate latitudes the rate of relative shift lies between 15° per hour and zero. This can be easily shown to equal 15° per hour multiplied by the sine of the latitude.

SUPPLEMENTARY READING

- H. A. Erikson, *Elements of Mechanics* (Chap. 8), McGraw-Hill, 1927.
John Perry, *Spinning Tops*, British Association Lecture, 1890, Sheldon Press, London, 1929.
H. Crabtree, *Spinning Tops and Gyroscopic Motion*, Longmans, Green, 1923.
A. M. Worthington, *Dynamics of Rotation*, Longmans, Green, 1925.

PROBLEMS

1. A light rectangular slab or lamina $ABCD$ measures 8 ft. along the AB edge and 6 ft. along the BC edge. It is loaded at the corners, in alphabetical sequence, with 5, 7, 3, and 9 lb. masses. Neglecting the weight of the lamina, calculate the system's moment of inertia about an axis normal to the surface of the lamina and through its center. *Ans.* 600 lb.-ft².

2. Calculate the moment of inertia, about a transverse axis through its center, of a disc whose radius is 20 cm, its density 9 g/cm^3 , and its thickness 12 cm. *Ans.* $27.13 \times 10^6 \text{ g-cm}^2$.

3. The mass of the earth is about 6×10^{21} tons and its radius about 4000 miles. Calculate its moment of inertia about its axis of rotation. *Ans.* $21.41 \times 10^{38} \text{ lb.-ft}^2$.

4. A rectangular parallelepiped rotates about an axis normal to the center of its ab surface. Its length a is 18 cm, its width b is 9 cm and its thickness is 4 cm. Its moment of inertia is found to be $1.35 \times 10^5 \text{ g-cm}^2$. What is its density? *Ans.* 6.17 g/cm^3 .

5. What is the moment of inertia about a transverse axis through its center of a ring of rectangular section, if its mass is 2 kg and its inner and outer radii are 20 cm and 22 cm respectively? *Ans.* $8.84 \times 10^5 \text{ g-cm}^2$.

6. If the ring described in Problem 5 is hung on a horizontal support as a hoop hangs on a nail in the wall, what is its moment of inertia about the support? *Ans.* $16.84 \times 10^5 \text{ g-cm}^2$.

7. Calculate the moment of inertia of the disc in Problem 2 about an axis at its edge and normal to its face. *Ans.* $81.39 \times 10^6 \text{ g-cm}^2$.

8. A thin rectangular lamina of metal is 80 cm long and rotates about an axis in its own plane but normal to its length and equidistant from its ends. If it weighs 3.6 kg, what is the moment of inertia, disregarding the thickness? *Ans.* $1.92 \times 10^6 \text{ g-cm}^2$.

9. Show that the moment of inertia of the lamina described in Problem 8 is four times as great if the axis is moved to one end and is still parallel to its first position.

10. An engine flywheel of 6 ft. radius and whose moment of inertia is 144,000 lb-ft², has a steady tangential force of 1500 lb. applied to its rim. What is the resulting angular acceleration? (take $g = 32$). *Ans.* 2 ra/sec^2 .

11. In Problem 10, calculate the final angular velocity, the number of revolutions made in 12 sec. from rest, and the work done. *Ans.* 24 ra/sec. ; 22.9 revolutions ; $13 \times 10^5 \text{ ft.-lb.}$

12. A solid disc of 20 cm radius rotates about an axis through its center and perpendicular to its plane faces. When it is revolving at 3 r.p.s. a tangential force of 2 kg is applied to its rim and stops it in 15 revolutions. Calculate the mass of the disc. *Ans.* 104 kg.

13. How long will it take the same force to stop the disc described in Problem 12? How long would it take if the disc were making 80 r.p.s.? *Ans.* 10 sec.; 267 sec.

14. A ball weighing 12 kg rolls in a bowling alley at a rate of 10 m per sec. What is the total kinetic energy? *Ans.* 840 joules.

15. How far would the ball in Problem 14 roll, if a force of a million dynes acting horizontally through its center opposes it? *Ans.* 84 m.

16. A ring whose mass may be regarded as being all equidistant from its center rolls down an inclined plane in 12 sec. How long would it take a solid disc to do the same? *Ans.* 10.38 sec.

17. A solid disc rolls down an inclined plane 200 cm long and 50 cm high. Calculate its final linear velocity and the time of descent. *Ans.* 256 cm/sec.; 1.57 sec.

18. A mass of 200 g stretches a helical spring supporting it through a distance of 6.8 cm. It is then set vibrating up and down. Calculate the period of vibration and the elastic constant of the spring. *Ans.* 0.52 sec.; 28,823 dynes/cm.

19. The torsion pendulum shown in Fig. 76 has a radius of 6 cm and a mass of 1.5 kg. A couple, consisting of two tangential forces of 40 g each, twists it through 60° . Calculate the coefficient of restitution and the period of vibration. *Ans.* 44.94×10^4 dyne-cm/radian; 1.54 sec.

*** 20.** A uniform bar 1 m long and 5 cm broad weighs 3.6 kg. Calculate its period when swinging about an axis 20 cm from one end. *Ans.* 1.52 sec.

*** 21.** Locate the terminus of the radius of gyration and the center of percussion in Problem 20. *Ans.* 41.65 cm and 57.85 cm measured from the axis.

CHAPTER 8

Elasticity

115. Meaning of elasticity: If the outlet of a tire pump is closed and the plunger is pushed in, one has to push harder and harder as the air is increasingly compressed. Then if the plunger is released and if there is no leak, it tends to spring back to its original position. The catgut "strings" of a violin are stretched in tuning them, the stretching force increasing as they lengthen. But when let down again, they recover practically their original length. Similarly a long steel wire clamped at one end may be twisted at the other, with a torque which increases with the twist. But it untwists again when released. These are all examples of a property of matter known as **elasticity**.

As long ago as 1676 the foundation of the theory of elasticity was laid in the publication of the famous law previously discovered by Hooke, which now bears his name. Hooke's law, as he worded it in Latin, may be rendered: "As the distortion, so the force," or more freely: "the change of form is proportional to the deforming force."

This is true, however, only for deformations which do not exceed some definite limit. This limit is different with different materials, and if it is exceeded, the deformation proceeds at an increasing rate as the action increases, and there is no complete recovery thereafter. This limit is known as the **elastic limit** of the material, and will be considered in more detail farther on.

116. Stress and strain. When a force acts upon a body, causing deformation, the internal reaction which tends to restore the original form is known as **stress**, and it is measured in terms of the *intensity* of the applied action.

The simplest intensity of action is force per unit area, or pressure (or its opposite, tension). It is measured in such units as dynes per square centimeter, or pounds per square inch, and its dimensions are

$$[p] = [MLT^{-2}] \times [L^{-2}] = [ML^{-1}T^{-2}].$$

Strain is the deformation associated with stress, and it is measured in terms of the *change* in some measure of the body, such as its volume or length, divided by the total measure. It is thus a ratio of two

like quantities and is a pure number without dimensions; it may be called **proportional deformation**.

As an illustration, suppose a force of a kilogram acts upon the upper surface of a cubical block whose edge is two centimeters. Then the area of a face is 4 cm^2 and the pressure or stress set up within the block is $250 \text{ g/cm}^2 = 250 \times 980 \text{ dynes/cm}^2$. If this stress should reduce the height of the block by one millimeter without altering its other dimensions, the change of volume is $0.1 \times 4 = 0.4 \text{ cm}^3$, and the strain is the ratio of this change to the total volume, or $0.4 \div 2^3 = 0.05$.

117. Modulus of elasticity. The ratio of stress to strain is known as the **modulus, or coefficient of elasticity**. Thus in the relation expressing Hooke's law,

$$\text{stress} = E \times \text{strain},$$

where E is the elastic modulus used here as a general term applying to any kind of elasticity. Let us consider the special case of a body whose volume is being changed by a force F acting on its surface. Then the change of pressure or stress, denoted by Δp , is given by $\Delta p = F/A$. If the total volume is V , and the change in V is denoted by ΔV , the strain is $\Delta V/V$. Therefore the volume or "bulk" modulus is given by

$$B = \frac{\text{stress}}{\text{strain}} = \frac{\Delta p}{\Delta V/V} = \frac{V\Delta p}{\Delta V}.$$

118. Young's modulus. This special coefficient relates only to wires, or other long cylinders of small radius, under tension. It recognizes only change in length, ignoring any change in section. If a force F is applied to one end of such a wire, the other being fixed, the stress reaction is F/A as above, where A is the sectional area of the wire. The strain is $\Delta l/l$, where Δl is the increase in length and l the mean total length. Therefore Young's modulus is found from

$$Y = \frac{Fl}{A\Delta l}.$$

If the tension is produced by a weight hung on the end of the wire, then

$$Y = \frac{mgl}{A\Delta l}.$$

119. Shear modulus. When a couple is applied to a body in a manner suggested by Fig. 81, it sets up an internal stress known as

a **shear**. This name is derived from the effect of a pair of shears whose edges tend to slide successive planes of the material over each other like a beveled pack of cards without altering the total volume. In computing the elastic modulus corresponding to this type of stress and strain, the intensity of the force, or the stress, is one of the forces of the couple divided by the area of the plane along which it acts, as shown in Fig. 82. The strain is the displacement s per unit length l . This ratio may be set equal to the angle θ , since s/l is usually very small. Consequently the shearing modulus, or modulus of rigidity, is given by $n = Fl/As = F/A\theta$. It can be shown that Young's modulus depends upon both n and the volume modulus

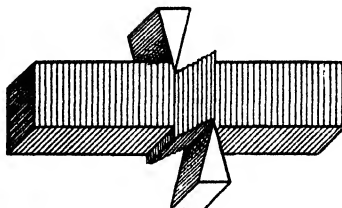


Fig. 81.

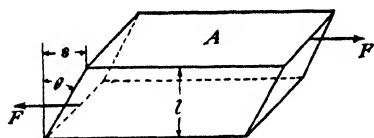


Fig. 82.

B , so that we may calculate Y from the equation $Y = B + 4n/3$.

120. Torsion. Let a cylindrical rod or wire, as shown in Fig. 83, be subjected to a torque L , caused by the couple Fd acting at one end while the other is held rigid. This

torque represents stress, and the resulting strain is the angular twist θ of the free end per unit length, or θ/l . The ratio T of stress to strain is given by $T = Ll/\theta$. Thus T is the torque required to twist a rod of unit length through one radian, and has been appropriately (though not usually) called the *torsion modulus* of the rod. If we divide T by l , we obtain

$$c = T/l = L/\theta,$$

where c is the torque required to rotate a rod of any length through one radian. It is known as the **moment of torsion**, and is the same quantity as the constant c used in connection with the torsion pendulum in Article 106.

It can be proved that $c = \pi r^4 n / 2l$, where n is the modulus of rigidity. It depends upon n , because when the rod is twisted, a shear is set up between its successive sections.

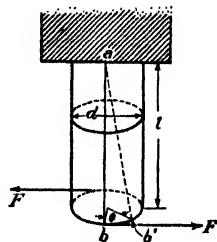


Fig. 83.

121. Elastic constants. The graph representing the relation between the stress of tension and strain is a straight line, as long as the

body being stretched retains the properties of a perfectly elastic material. This is because their ratio is constant. But at some point *A*, as seen in Fig. 84, the *elastic limit* is reached, the strain begins

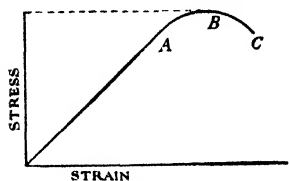


Fig. 84.

to increase more rapidly than before, and the recovery from beyond this point is no longer complete. Finally at *B*, known as the *yield point*, the stress which the body is able to support reaches a maximum and thereafter it resists less and less, though with rapidly increasing strain, until at *C* it yields completely through fracture.

122. Impact. If a mass m moving with a velocity v strikes another body at rest, the resulting motions may be calculated with ease provided the elasticity of the two bodies is perfect. In such cases the problem can be solved by the principle of the conservation of energy and of the conservation of momentum.

First let us suppose that one of the bodies is immovable, as is the case when a billiard ball strikes an ideal cushion which has perfect elasticity and therefore absorbs no energy. If the original direction is normal to the cushion, the recoil will be normal also, and in accordance with the principle of the conservation of energy, $mv_1^2/2 = mv_2^2/2$ and $v_1 = v_2$, or the velocities before and after impact are equal, though oppositely directed.

The *average force* of impact can be determined only if the time or distance it takes to stop the moving body is known. If the time is known, $Ft = mv$ gives the value of F . If the distance to which the moving body penetrates a stationary body is known, then the equation $Fs = \frac{1}{2}mv^2$ makes the calculation possible. In general, however, the problem is insoluble. In a certain widely quoted examination, the candidate was asked to calculate the force with which a certain object dropped from a given height would strike the top of a table. This is unanswerable. If the table were of iron, the force would be much greater than if it were made of wood, or some still softer material. A hard surface stops a moving object so quickly that the force of impact is often great enough to break it. A fragile object dropped upon a stone floor is more likely to break than if the floor were of wood.

123. Collision. When both bodies are in motion, the resulting impact, or collision, is subject to precisely the same laws as when one is stationary, for it is the velocity with which the two bodies meet each other that is important, and not whether one or the other is at

rest, rest being a purely relative term.† Let u_1 and v_1 be the velocities before and after collision of a body of mass m_1 with another body of mass m_2 , moving in the same straight line. Let u_2 and v_2 be the velocities of m_2 before and after the collision. Equating initial and final momenta, we have

$$m_1u_1 + m_2u_2 = m_1v_1 + m_2v_2. \quad (1)$$

Evidently there are now two unknown quantities (assuming the data before collision are given), and a second equation is necessary for their solution. This is supplied by a law discovered by Newton, according to which the velocity of separation after collision is always proportional to the velocity of approach. The constant of proportionality depends upon the nature of the bodies, and is called the coefficient of restitution, being expressed by

$$e = \frac{v_2 - v_1}{u_1 - u_2}. \quad (2)$$

If the bodies are perfectly elastic, e is unity, and the velocities of approach ($u_1 - u_2$) and separation ($v_2 - v_1$) are the same. In general, however, e is less than unity, and $(v_2 - v_1) < (u_1 - u_2)$.

124. Problems concerning impact. In such problems involving motions in one straight line, great care must be observed to use the correct sign for each velocity, according to whether it moves in a positive direction (to the right), or in a negative direction (to the left).

The velocities after impact are usually the unknown quantities. Each is obtained by eliminating the other from equations (1) and (2) above, giving

$$v_1 = \frac{u_1(m_1 - em_2) + m_2u_2(1 + e)}{m_1 + m_2},$$

and

$$v_2 = \frac{u_2(m_2 - em_1) + m_1u_1(1 + e)}{m_1 + m_2}.$$

For the special case where the colliding bodies have the same mass and when $e = 1$, by setting $m_1 = m_2$ we have

$$v_1 = \frac{2m_2u_2}{2m_2} = u_2, \text{ and } v_2 = \frac{2m_1u_1}{2m_1} = u_1,$$

which shows that the bodies have only exchanged velocities.

When one of the bodies is fixed and $e = 1$, then $u_2 = v_2 = 0$, and from equation (2) of Article 123, $-v_1/u_1 = 1$. Therefore the moving body rebounds, with its original velocity exactly reversed.

† In the following discussion, and in the problems based upon it, we shall consider only "head-on" collisions and disregard the possibility of rotation, as this would reduce the momentum of translation and greatly complicate the problem.

In this last case, if e is less than unity, since $v_1 = -eu_1$, the velocity after impact is less than it was before. Thus if the moving mass falls from a height H on a horizontal surface, its velocity of impact, according to the laws of falling bodies, is $u_1 = \sqrt{2gH}$. If h is the height to which it rebounds, then $v_1 = \sqrt{2gh}$. Therefore, substituting in $v_1 = -eu_1$, we obtain

$$e = \sqrt{\frac{-h}{H}},$$

where the negative sign shows that h and H are measured in opposite directions.

The fact that e can be expressed in terms of the ratio of the height of rebound to the distance fallen through is made use of in an instrument known as the scleroscope. It is designed to test the hardness of metallic surfaces. A very hard steel ball, or a plunger with a diamond end, is dropped from a known height, and the rebound carefully measured. This makes a basis of comparison of the values of e as existing between the same ball and the various surfaces tested, thus giving a measure of their relative "hardness."

Table of Elastic Constants †

Substance	Young's Modulus, Y (dynes per cm ²)	Rigidity Modulus (Twisting Shear) n (dynes per cm ²)	Volume Modulus, B (dynes per cm ²)	Compressibility, $1/B$ (cm ² per dyne)
Copper.....	12.4 to 12.9 × 10 ¹¹	3.9 to 4 × 10 ¹¹	14.3 × 10 ¹¹	0.70 × 10 ⁻¹²
Iron				
(wrought)	19 to 20 "	7.7 to 8.3 "	14.6 "	0.68 "
Iron (cast) ..	10 to 13 "	3.5 to 5.3 "	9.5 "	1.04 "
Steel.....	19.5 to 20.6 "	7.9 to 8.9 "	18.1 "	0.55 "
Brass.....	9.7 to 10.2 "	3.5 "	10.65 "	0.94 "
Glass (flint) ..	5 to 6 "	2.0 to 2.5 "	3.6 to 3.8 × 10 ¹¹	2.6 to 2.8 × 10 ⁻¹²
Water.....	—	—	0.205 × 10 ¹¹	48.9 × 10 ⁻¹²
Mercury.....	—	—	2.6 "	3.82 "
Ether.....	—	—	0.07 "	145.2 "

† The very large numbers needed to express elastic moduli are partly due to the fact that stress is measured by a very small unit (dyne/cm²), and partly because really great forces are needed to produce appreciable strains in solid bodies. For instance a pressure of 1400 atmospheres (see Article 129) is needed to produce a volume diminution of one-tenth per cent in copper.

SUPPLEMENTARY READING

H. A. Erikson, *Elements of Mechanics* (Chap. 12), McGraw-Hill, 1927.
M. & T. Merriman, *Strength of Materials*, Wiley, 1928.

PROBLEMS

1. A copper wire 2 m long and 0.5 mm in diameter supports a mass of 3 kg. It is stretched 2.38 mm. Calculate Young's modulus. *Ans.* 12.6×10^{11} dynes/cm².

2. A force of 100 kg is exerted on a piston sliding in a tube filled with water. The column of water compressed by the piston is 2 m long and 1 cm in diameter. How far does the piston move in compressing the water? *Ans.* 1.22 cm.

* 3. Forces of 800 kg each are applied to a wrought-iron rod as in Fig. 83. The length of the rod is 6 m, its diameter 4 cm, and its modulus of rigidity 8×10^{11} dyne/cm². What is the displacement of a point *b* at the lower end? *Ans.* 1.87 mm.

4. A strip of silk cloth measuring 15×40 cm, when pulled lengthwise by a force of 4 kg, stretches 3 mm. Calculate its elastic modulus of surface deformation. (NOTE: The stress is measured in dynes per linear cm.) *Ans.* 3.48×10^7 dynes/cm.

5. A metal ball dropped on a wooden table from a height of 9 ft. rebounds to a height of 4 ft. Calculate the coefficient of restitution. *Ans.* $\frac{2}{3}$.

6. Two bodies whose masses are 4 lb. and 12 lb. are moving in the same direction with velocities of 12 ft./sec. and 3 ft./sec. respectively. The lighter body overtakes the heavier one. What are the final velocities if the coefficient of restitution is $\frac{1}{3}$? *Ans.* 3 ft./sec. and 6 ft./sec. in the same direction.

7. Two bodies whose masses are 30 and 80 g are moving in opposite directions with velocities of -6 and $+4$ m/sec. respectively. What are their velocities after collision if $e = 0.4$? *Ans.* 418.2 cm/sec., and 18.2 cm/sec.

* 8. How much energy is lost in the collision of Problem 7? (NOTE: Calculate the total kinetic energy before and after the collision.) *Ans.* 0.92 joule.

* 9. A moving body strikes a free and stationary body of twice its mass, and continues in the same direction with a velocity of 80 cm/sec. If $e = 0.4$, what velocity is imparted to the body struck? *Ans.* 560 cm/sec.

CHAPTER 9

Hydrostatics

125. Fluids. Most people would say that a fluid is anything that flows. That is of course true, but lead is not a fluid, and yet under sufficient pressure lead may be made to flow through a small hole. The fact is that most things can be made to flow, so that to distinguish solids from fluids we must use some other characteristic. Suppose we try shape as a basis for definition. Fluids certainly have no shape, while solids have one, and they resist any attempt to deform them. We may then say that solids have rigidity of form, and fluids do not. But even this definition is not perfect, because some viscous fluids approach the behavior of very soft solids, and can have temporary form apart from any container. However, in general, form is characteristic of solids and formlessness is characteristic of liquids.

The reason why fluids have little or no rigidity of form is that their molecules have great freedom of motion, while in solids the motion of the molecules is restricted to a very limited region about a mean position of rest.

126. Liquids and gases. A liquid is a fluid which possesses a definite volume at a definite temperature, while a gas is a fluid which can occupy any volume by expanding or being compressed until it just fills it. This means that liquids resist a change in volume just like solids, and may have even greater bulk elasticity. Gases also resist compression, but they tend to expand indefinitely when the pressure is withdrawn.

The inner structure which accounts for these differences is only partly understood, but at any rate the molecules of a liquid, though free to move among each other, are held together by intimate bonds which keep their average distances from each other constant at constant temperature. In gases these bonds are extremely weak, and a molecule moves with almost perfect freedom in a straight line between collisions with other molecules, or with the walls of the containing vessel. The average distance between collisions is called the **mean free path**, and it increases as the gas expands to occupy larger volumes under reduced pressure.

127. Compressibility. This is the reciprocal of the modulus of volume elasticity; therefore it is the ratio of the proportional change in volume to the pressure. This change of volume in solids and liquids is extremely small. To produce measurable changes the pressure must be large and is usually measured in atmospheres.

Both volume elasticity and compressibility are constant only within certain limits, and it is therefore not unusual to express them in terms of infinitesimal changes of pressure and volume while the total remains substantially unaltered. Then $E = dp/(dv/v) = v(dp/dv)$ and the compressibility is dv/vdp . These expressions are also used with gases, and play an important part in the theory of thermodynamics.

128. Cohesion and viscosity. The bonds just mentioned between the molecules of a liquid are due to a force known as **cohesion**, but they must be very close together to experience this force, and it is therefore practically absent from gases. Solids of course have vastly more cohesion than liquids, owing to the relative immobility of their molecules. The tensile strength of a bar of iron is due to forces of cohesion between its molecules.

Adhesion is the same type of force as cohesion, but the name is reserved for the attraction between molecules of *different* substances, as in the adherence of water to glass.

Viscosity is fluid friction which opposes the sliding of one layer of a fluid over another, as if the layers were separate sheets of some rough material. If we set the upper layers of a glass of water rotating around the axis of the glass, it will not be long before all the water will be rotating in the same manner. Gases also possess this property, as can be shown by spinning a disc about a vertical axis over another disc lightly pivoted and just below it. The second disc soon begins to turn also, showing that the air near the upper one has been set going due to collisions of its molecules with the rotating surface, and that this motion has been passed on downward to the second disc.

This transfer of motion from a rapidly moving layer of a gas to one in slower motion is known as a **transport phenomenon**. It is brought about by the transfer of the forward momentum of the molecules in excess of that possessed by the adjacent and more slowly moving layer. This excess momentum is passed along from layer to layer by means of collisions between the molecules, and is possible only because they have some freedom of motion and a finite diameter.

129. Pressure on fluids. (A) Since fluids cannot support a stress due to forces which tend to change their shape, the only stress possible when a fluid is at rest is one which resists change of volume.

This is the reaction against an applied pressure, or force per unit area. *It acts everywhere normal to the fluid's surface*, and since action is equal and opposite to reaction, pressure may be thought of either as applied to, or existing within, the fluid.

Fluid pressure is usually measured in dynes per square centimeter (whose unit is known as a **bar** or **barye**), in pounds per square inch, or in atmospheres. This latter unit is the atmospheric pressure when the barometer stands at an arbitrarily assumed average height. In the c.g.s. system this is 76 cm and in the English system 30 inches, which is not quite the same. The former corresponds to a pressure of 1,013,300 bars, roughly a million. A million bars is called a **mega-bar**, but in European practice this is erroneously called a bar, so a kilobar in America is the same as a millibar in Europe. The barometric height of 30 in. corresponds to a pressure of 14.736 lb./in²., although 76 cm indicates 14.697 lb./in².

Still another way of denoting pressure is in *millimeters* or *inches of mercury*. This is the height of the barometric column under the pressure considered. Thus 38 cm of mercury is half an atmosphere, 6 in. is one fifth, and their values in standard units can be found from the figures given above. But it should be remembered that unless a problem involves a ratio of two pressures (when any unit is allowable) we cannot use arbitrary units such as atmospheres or "millimeters of mercury." If these are given they must be translated into true pressure units of force per unit area.

130. Pressure within fluids. (B) Since true fluids are highly mobile bodies, *any pressure communicated to the surface of a confined fluid is transmitted unchanged to every part of the interior*. This is known as *Pascal's principle*, and was established by him after a series of experiments on liquids during the middle of the seventeenth century.

(C) *The internal pressure is equal in all directions at any point*. This may be shown by direct experiment, or inferred from the fact

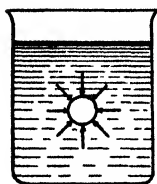


Fig. 85.

that if a *very small* solid sphere of the same density as the fluid is wholly immersed, it must remain at rest. Then the vector sum, or resultant, of all the forces acting normal to the sphere must be zero, and they must therefore all be equal, as indicated in Fig. 85. The two principles (B) and (C) stated above, as well as (A) (pressures are normal to bounding surfaces) are all applicable to both gases and liquids. But there are

additional principles applicable to liquids alone which result from their nearly perfect incompressibility and from the possibility of a

free surface. Gases also have special laws, resulting from their tendency to indefinite expansion, which are inapplicable to liquids.

131. Pressure within liquids. If the pressures at varying depths in a liquid are measured by any convenient pressure gauge, it appears that if the liquid is practically incompressible, and if gravity alone is acting on it, then the pressure varies directly as the depth and the density of the liquid. This is readily proved to be a necessary consequence of Pascal's principle, because the downward force on a horizontal area a at a depth h may be regarded as due to the weight of the vertical cylinder of the liquid resting upon it as shown in Fig. 86. The volume is then ah , and if the density is d , the mass is ahd . The force exerted by this mass is $ahdg$. This force divided by the area gives the pressure, or

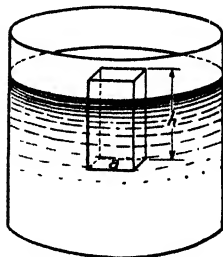


Fig. 86.

$$p = hdg.$$

According to Pascal's principle this pressure is equal in all directions at the depth h . This important relation is the basis of all hydrostatic problems, when no other force than gravity is acting on the liquid.†

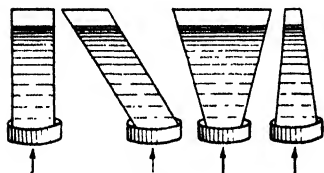


Fig. 87.

One very important consequence is that the total force on the horizontal bottom of a vessel of any shape depends only upon the depth of the liquid and the area of the base. This was experimentally established by Pascal, who used vases like those in Fig. 87, all having the same sized base, and found that the forces required to support the cap which closed the bottom were all equal when the vases were filled to the same level.

132. Force on submerged surfaces. Let a submerged plane of area A , shown in Fig. 88, be divided up into infinitesimal areas, $a_1, a_2, a_3, \dots, a_n$, being the bases of cylindrical columns of the liquid at depths $y_1, y_2, y_3, \dots, y_n$ below the

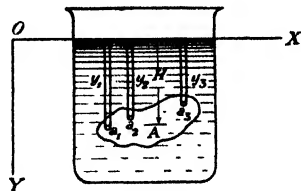


Fig. 88.

† Obviously, hdg gives the pressure in absolute units. If p is to be expressed in gravitational units, such as lb./in^2 , we must divide hdg by k , as explained in Article 50. Then $p = hd/k$, since $k = g$ numerically.

surface. The total force on A due to all the y columns is given by

$$F = \Sigma aydg,$$

or

$$F = dg\Sigma ay, \quad (1)$$

since dg is constant. But the center of area of a plane surface is obtained from the equations

$$X_c = \frac{\Sigma ax}{\Sigma a}, \text{ and } Y_c = \frac{\Sigma ay}{\Sigma a}. \quad (2)$$

Therefore the depth Y_c , or H , as indicated in Fig. 88, is given by

$$H = \frac{\Sigma ay}{\Sigma a}, \text{ or } H\Sigma a = \Sigma ay. \quad (3)$$

Then setting $A = \Sigma a$, the total area, we may substitute AH for Σay in equation (1) and so obtain

$$F = AHdg. \quad (4)$$

This important equation of hydrostatics gives the total force on a submerged inclined plane in terms of the depth H of the center of area, or *center of pressure*. The average pressure over the surface is obtained by dividing by the area, and is given by

$$p_{av} = Hdg. \quad (5)$$

If the plane is rectangular or a parallelogram, the center of area is the intersection of the diagonals. If it is triangular, it is on the intersection of the median lines and therefore two thirds the distance from a vertex to the side opposite. Thus for such simple figures H can be found, if the exact position and slope of the surface are known.

133. The hydrostatic paradox. In a liquid of given density, dg is constant, and F then depends only on the product AH . It is there-

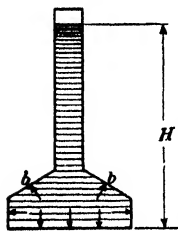


Fig. 89.

fore possible to exert enormous forces with a very small amount of liquid acted on by gravity alone. Thus in Fig. 89, suppose a liter of water fills the vessel to a height of a meter, and that the area of the base A is 1000 cm^2 ; then the force on the base is equal to $AHdg = 1000 \times 100 \times 1 \times 980 = 98 \times 10^6$ dynes, or $F = 100$ kilograms, produced by one kilogram of water. It is well to note that in problems where the *total* pressure, or *total* force at a given depth in a liquid is desired, the atmospheric pressure at the

surface must be included. However, the outer surface of the containing vessel is subject to the same pressure; therefore the net force on the base in the preceding problem is due to the liquid alone.

Further, the upward forces on the container, indicated by $b\ b$, tend to neutralize the downward force just calculated, so that if placed on a balance the total weight of the apparatus would be only that of water and vessel combined.

134. Communicating vessels. If two vessels of any shape, connected by any kind of tube, are filled with a liquid in such a way that the tube is completely filled also, then the liquid stands at the same level in each. This may be shown to follow from the principles already explained. At a point P in the tube (Fig. 90), the opposing pressures p and p' , tending to cause the liquid to flow, are equal and opposite, and there is therefore no motion. The pressure p is due entirely to H , because the two columns of height h below P are equal and balance each other. Similarly to the right of P , the pressure p' is due to H' only, because the two h' columns are balanced. Therefore $p = Hdg = p' = H'dg$, and $H = H'$. This principle is often expressed by the familiar saying: "Water seeks its own level."

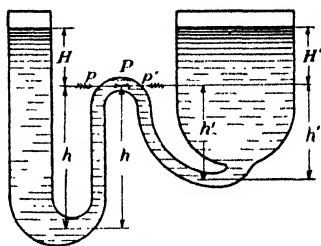


Fig. 90.

135. The siphon. This useful device, which enables us to draw a liquid out of a container without tilting it, may be explained in a similar manner, though in this case there is no equilibrium. If the bent tube shown in Fig. 91 is completely filled with the liquid, the atmospheric pressure P acting on the free surface at L is transmitted unchanged through the tube, and acts as a pressure tending to cause motion to the right at the point a . It is however diminished by the downward and unbalanced pressure due to the column of height h , or hdg . Similarly the atmospheric pressure acting at b is transmitted unchanged to a , where it tends to cause motion in the opposite direction, but this in turn is diminished by the pressure Hdg due to the liquid in the longer leg of the siphon. The net pressure which causes the flow is the algebraic sum of the four preceding items, or

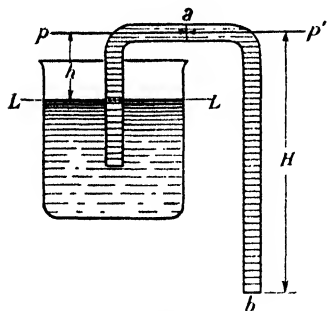


Fig. 91.

$$p - p' = P - hdg - (P - Hdg) = Hdg - hdg.$$

$$\therefore p - p' = (H - h)dg.$$

Evidently b must be lower than L if an outward flow is to occur, but if $H = h$, there is equilibrium, and when $H < h$, the liquid in the tube flows back into the container.

136. The barometer. If the vertical tube closed at the upper end, shown in Fig. 92, is filled with a liquid, and if the open end is immersed

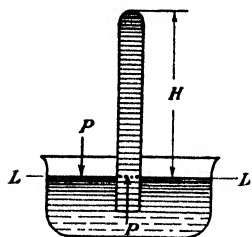


Fig. 92.

in a vessel containing the same liquid, the atmosphere acting on the free surface causes an equal pressure P to act within the tube at the same level. This upward pressure is opposed by that due to the column of height H , or $p = Hdg$, so that if p is less than P , the tube remains filled. But if the tube is very long like the first tube in Fig. 93, the atmospheric pressure may be less than enough to support a column which fills it. In this case,

since $P = Hdg$, the liquid is sustained at a height $H = P/dg$, and above the level MM the space is filled only with the vapor of the liquid. Such a "vacuum" above the barometric column was first investigated by the Italian philosopher Torricelli, and is named after him, a *Torricelli vacuum*. The height of a column of mercury which is supported by standard atmospheric pressure is 76 cm, as has already been stated. But if the column is of water, it stands much higher, because water is much less dense than mercury. This height under standard conditions may be determined as follows: Since $H = P/dg$, it is clear that at constant pressure the barometric height varies inversely as the density of the liquid. We may then calculate the barometric height H' of a column of water from the proportion $H'/H = d/d'$. The density of mercury d is 13.596 at 0° , and $H = 76$ cm, while the density d' of water is almost unity. Therefore $H' = 13.596 \times 76 = 1033$ cm, or 33.89 ft. at that temperature.

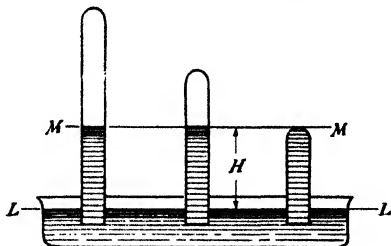


Fig. 93.

The height of the mercurial column serves as a valuable means of measuring atmospheric pressure under varying conditions. At a given altitude a "fall" indicates a low pressure area often associated with storms, while a *gradual* rise is generally associated with fair weather.

As atmospheric pressure is caused by the weight of the air above us, just as hydrostatic pressure is due to the weight of a liquid, the

pressure diminishes as we rise in altitude, because there is then less air to produce it. The decrease of pressure with altitude is not uniform as we ascend, as it would be under water, because the density of the air diminishes rapidly under reduced pressure. There is therefore no wholly satisfactory way of calculating altitude in terms of pressure as measured by the barometer.

137. The lift pump. The height of the liquid in a barometric tube has been shown to be caused by the atmospheric pressure acting on its free surface, with no counterpressure except that of the liquid's vapor acting within. If the top of the tube is not closed, but fitted with a piston as shown in Fig. 94, the atmosphere acting on the free surface will cause the liquid to follow the rising piston until it has reached the critical height $H = P/dg$ of the barometric column. The liquid will then remain at that level and any further rise of the piston creates only a Torricelli vacuum.

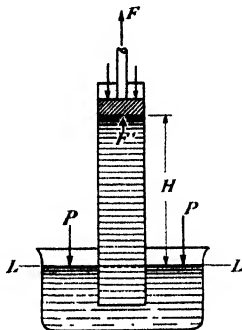


Fig. 94.

The force F required to lift the piston when it is at height h is equal to the downward force PA exerted by the atmosphere on its upper surface, less the upward force $F' = PA - Ahdg$, where A is the cross section area of the tube. Therefore

$$F = PA - (PA - Ahdg) = Ahdg,$$

which becomes equal to PA when $h = H$. Thus we have to exert a greater and greater force until the column reaches the barometric height. After that, F is constant and equal to PA .

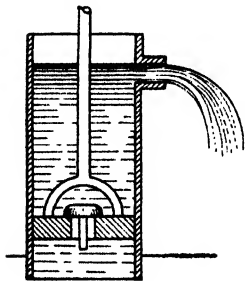


Fig. 95.

In actual water pumps of this type, owing to leakage around the piston, or "bucket," as it is called, and to air mixed with the water, the ideal height of 33.9 ft. is never achieved, and 28 ft. is regarded as a fair performance. In order to deliver the water thus raised by the atmospheric pressure to a higher level, a valve known as the bucket valve is used, as shown in Fig. 95. It opens during the downstroke of the piston and closes under the pressure of the water above it, on the upstroke. Thus it lifts the water trapped above it to any reasonable height.

138. The hydraulic press. This is a true machine in the sense that it transforms a small force into a large one at the expense of motion, the work input being equal to the work output, if friction is neglected. The object to be compressed is placed between the movable and stationary plates D and D' , as shown in Fig. 96. The two pistons B and C slide in communicating cylinders filled with some liquid, usually oil. The valve V admits this liquid from a reservoir through the tube T , and the valve V' , by preventing its return from the large cylinder, locks the lower plate at the end of each stroke. Let a force f be applied to the piston B ;

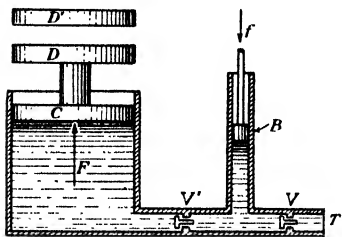


Fig. 96.

then by Pascal's principle the pressure p thus created appears unchanged throughout the liquid. Its value is f/a , and it exerts a force F on the large piston C equal to pA , or fA/a . Thus f is multiplied by a factor equal to the ratio of the areas of the two pistons, and this factor is the mechanical advantage of the press.

139. Archimedes' principle. The cave men must have known that wood floats on water, and that even a stone seems to weigh less under water than in the air. But apparently no one generalized these facts into the form of a law until about 250 B.C. when the great Greek philosopher Archimedes derived from observation the following principle which bears his name: *All bodies floating on or submerged in a fluid are buoyed up by a force exactly equal to the weight of the fluid they displace.* No actual test is necessary to obtain this principle, which may be rigorously derived from a purely imaginary experiment. Suppose that a portion of some liquid becomes solidified, as indicated by the irregular contour in Fig. 97. Suppose also that it retains its original density unchanged. Under these conditions all external forces acting on it, including gravity, are unchanged, and it must remain at rest as when it was still fluid. It is therefore supported by an upward force equal to its weight. Expressed in dynes or poundals this is given by vdg , where v is its volume and d is the density of both solid and liquid. Obviously any solid displaces its own volume of a fluid in which it is wholly submerged. We do not need Archimedes to tell us that! So vd is the mass of the displaced liquid and vdg

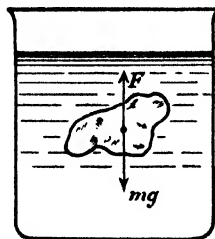


Fig. 97.

is its weight. This proves the famous principle for complete submersion.

Of course the same upward force acts no matter how dense is the solid which fills the volume v . If it is more dense than the liquid, its weight $vd'g$ is greater than the upward force vdg and it sinks. If it is less dense, the upward force is the greater, and it rises, until at the surface it displaces only a sufficient volume of the liquid to balance its own weight. In this case the apparent loss of weight is equal to $v'dg$, where v' is the volume of the liquid displaced, which is less than v , the volume of the body. In both cases, then, the buoyant force depends only upon the volume displaced and the density of the displaced fluid.

140. Weighing submerged bodies.

If the body has a density d' , greater than d , and is weighed both in air and in a liquid, as shown in Fig. 98, its weight when it is submerged is decreased by the buoyant force vdg . This apparent loss of weight may be expressed by the relation

$$w_a - w_s = vdg, \quad (1)$$

where w_a is the weight of the body in air, and w_s is its weight submerged. But

$$w_a = vd'g. \quad (2)$$

Therefore, dividing (1) by (2), we have

$$\frac{w_a - w_s}{w_a} = \frac{vdg}{vd'g},$$

$$\therefore w_a - w_s = w_a \frac{d}{d'}, \quad (3)$$

which gives the apparent loss of weight in terms of the two densities and the weight of the body in air.

141. Inverse of Archimedes' principle. Since a liquid exerts an upward force on a submerged body, the body must react upon the liquid with an equal downward force, according to Newton's third law. Thus if a stone is lowered by a string into a pail of water, the pail weighs as much more as the tension on the string is less, due to the buoyancy of the liquid. This is easily tested by putting a pail of water on a balance and observing the increased weight when the stone is lowered into it. The same is true if the stone is replaced by a rod or any other object held in the hand.

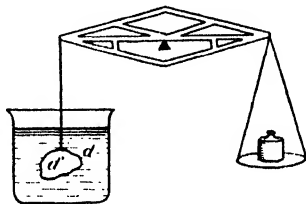


Fig. 98.

142. Specific gravity. The ratio of the density of any body to that of water is called its **specific gravity**, usually designated by the letter s . If the density of water were exactly unity, density and specific gravity would be numerically equal in the c.g.s. system. But this is not quite the case, and water's actual density must be taken into account if high precision is desired. In the English system the two quantities are very different, because the density of water is 62.4 lb. per cubic foot, while its specific gravity (compared to itself) is 1 in both systems. A further distinction between d and s is in their dimensions. Density is given by $[d] = [ML^{-3}]$, as has been stated, but specific gravity, being the ratio of two densities, is a pure number of no dimensions.

A convenient method for measuring the specific gravity of solids depends upon Archimedes' principle. If a body is submerged in water, we may obtain s from the apparent loss of weight, using equation (3), Article 140. This may be written

$$s = \frac{d'}{d} = \frac{w_a}{w_a - w_s}. \quad (1)$$

To determine s the body is first weighed in air to find w_a , and then it is submerged in water and weighed again. The difference of these quantities, $w_a - w_s$, gives us the buoyant force exerted by the water, and the specific gravity equals w_a divided by this force. If density instead of specific gravity is required, we must know the density of the water exactly, and then from the same equation

$$d' = d \frac{w_a}{w_a - w_s}, \quad (2)$$

where d is the density of water, or of any other liquid in which the body is submerged.

In these equations the weight has been expressed in absolute units given by mg . But it is unnecessary to introduce g into the final calculation, because $w_a/(w_a - w_s)$ is a ratio, so we may use any force unit such as the gravitational force-pound or force-gram.

143. Flotation. If a body is less dense than water when wholly submerged, it displaces a weight greater than its own, and if free to move, rises at once to the surface. Equilibrium is attained when just enough of the body remains under water to displace its own weight. This is expressed by boat builders as the **displacement** of a vessel, so that a vessel's displacement and actual tonnage are synonymous. Registered tonnage, as used in describing ocean steamers, is quite a

different measure, and refers more to their volume or capacity for cargo than to their actual weight.

If a body has the same density as water, it will neither float nor sink, but remain wherever it happens to be, as if it were water itself. But if it is only very slightly denser than water it may sink to the bottom of the deepest ocean. Whether it does so or not depends upon whether it is more, or less compressible than the surrounding medium. If more so, its relative density increases and it sinks faster and faster. If less so, it may reach a level where the water becomes sufficiently dense to support it.

144. Flotation problems. If it is desired to find what proportion of the volume of a floating body is submerged, the calculation is as follows: Let v_1 and v_2 be the volumes which are below and above the surface respectively, and let d' and d be the densities of the body and supporting liquid. Then the weight of the body is $(v_1 + v_2)d'g$, and it displaces a weight of liquid equal to v_1dg . But these are equal, and

$$v_1dg = (v_1 + v_2)d'g.$$

$$\therefore \frac{v_1}{v_1 + v_2} = \frac{d'}{d},$$

which is the ratio sought. As an illustration, an iceberg has a density of 0.9167 g/cm^3 , and if it floats in sea water of density 1.025 g/cm^3 , the proportion submerged is

$$\frac{d'}{d} = \frac{0.9167}{1.025} = 0.894,$$

so that nearly nine tenths of the iceberg is under water.

145. Stability of flotation. As in the case of objects resting on a plane surface, the question of stability depends upon the moment of the force of gravity, which must supply a restoring torque when equilibrium is disturbed. In flotation, however, the center of mass is not the only essential point to be determined. We must also know where the resultant of the fluid pressure may be regarded as acting. This point is the **center of buoyancy**, and is actually the center of mass of the displaced liquid. The upward thrust through the center of buoyancy of a vessel, shown at B in Fig. 99, combined with the force of gravity acting downward

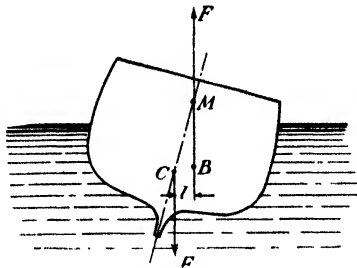


Fig. 99.

through C , forms a couple Fl which tends to rotate the vessel back to "even keel." If the line of action of the upward force through B is

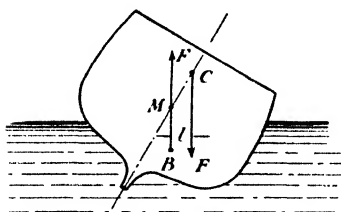


Fig. 100.

produced to intersect the ship's vertical axis of symmetry, the point M thus determined is known as the **meta-center**. As the ship heels over on her side, M shifts in position and l changes in length, but as long as M is above C , and l is greater than zero, there is a restoring couple, and the vessel rights herself. But if she has

been overloaded with a deck cargo, or badly designed, so that C is too high, the situation in Fig. 100 arises. The metacenter is now below C , and the couple Fl tends to capsize the ship, with l increasing as her "list" increases.

SUPPLEMENTARY READING

King and Wisler, *Hydraulics* (First five chapters), Wiley, 1927.

Table of Densities (g/cm³)

Substance	Density at 20°C.	Substance	Density at 20°C.
Aluminum.....	2.65	White Pine Wood..	0.4 to 0.5
Copper.....	8.93	Cork.....	0.22 to 0.26
Gold.....	19.32	Common Glass.....	2.5
Iridium.....	22.41	Paraffin Oil.....	0.8
Iron (pure).....	7.86	Glycerine.....	1.26
Lead.....	11.37	Gasoline.....	0.68 to 0.72
Mercury.....	13.546	Alcohol (ethyl).....	0.79
Platinum.....	21.50	Sulphuric Acid (concentrated)...	1.84
Silver.....	10.5	Air (0°, normal atmosphere).....	1.293×10^{-3}
Zinc.....	7.1	Hydrogen ".....	0.0899×10^{-3}
Brass.....	8.4 to 8.7	Oxygen ".....	1.429×10^{-3}

PROBLEMS

1. Calculate the pressure in dynes/cm² (or bars), in mm of mercury, atmospheres and in lb./in². at a depth of 15 m in a liquid whose density is 2.4 g/cm³. *Ans.* 3,528,000 bars; 2646 mm; 3.48 atmospheres; 51.17 lb./in².

2. What is the total force on a submerged rectangular area 12 × 16 cm when it is inclined at 30° to the horizontal and its upper edge of 12 cm is 20 cm below the surface of a jar of water? *Ans.* 4.5 megadynes.

3. A dam blocks a V-shaped trough filled with water. The water stands 15 ft. above the bottom of the V, and the width of the trough at this height is 12 ft. What is the total force on the dam? (A cu. ft. of water weighs 62.4 lb. nearly.) *Ans.* 28,080 lb.

4. Calculate the total force on the sides and base of a rectangular box whose base measures 6×4 ft., and which contains water to a depth of 5 ft. *Ans.* 23,088 lb.

5. Taking the density of mercury as 13.6 g/cm^3 , calculate in feet the height of a column of water which is equivalent to a barometric height of 30 in. of mercury. *Ans.* 34 ft.

6. A vertical cylinder of 24 cm internal diameter is filled with water. A frictionless plunger whose diameter is 4 cm is pushed in through a bushing in the upper head with a force of 2 kg. This forces down a piston just fitting the cylinder at the lower end. The distance between plunger and piston is 60 cm. What force is exerted on the piston? *Ans.* 99.13 kg.

7. A stone weighs 140 g in air, and 81.66 g when submerged in water. What is its specific gravity? *Ans.* 2.4.

8. A piece of copper whose density is 8.93 g/cm^3 weighs 180 g in air and 162 g when submerged in a certain liquid. What is the density of the liquid? *Ans.* 0.893 g/cm^3 .

9. A piece of glass of unknown density loses 43.71 g when weighed in water and 80.36 g when weighed in concentrated sulphuric acid. What is the specific gravity of the acid? *Ans.* 1.84.

10. A block of wood of rectangular section and 6 cm deep floats in water. If its density is 0.6 g/cm^3 , how far below the surface is its lower face? *Ans.* 3.6 cm.

11. What weight placed on the upper surface of the block in Problem 10 is needed to sink it to a depth of 5 cm, if its area is 120 cm^2 ? *Ans.* 168 g.

12. A cylindrical block of wood of 5 cm radius and 20 cm long is loaded at one end with a mass of iron weighing 800 g. How much does it project above the surface of the water? (The density of the wood is 0.4 g/cm^3 , and of the iron 7.8 g/cm^3 .) *Ans.* 3.12 cm.

13. If the liquid in Problem 12 has a density of 1.4 g/cm^3 , how far does the cylinder project? *Ans.* 8.31 cm.

CHAPTER 10

Mechanics of Gases

146. Boyle's law. When an automobile tire is inflated, the pump compresses the air into a smaller volume and raises its pressure from about fifteen pounds per square inch to thirty pounds or so. If the tire is punctured, the air rushes out, often with explosive violence, and its pressure again falls to that of the surrounding atmosphere. In this way, air and other gases behave very differently from liquids, which have a nearly constant volume under all pressures.

The relation between the pressure and volume of gases was first discovered by the English physicist, Robert Boyle, in 1662, and independently by Mariotte in France a few years later. According to this law, *the pressure of a gas varies inversely as its volume at constant temperature*, or $pv = b$, where b is a constant. This is rigorously true only for what may be called an ideal gas. Real gases deviate more or less from this law, according to how near they are to liquefaction, when they cease to be gases at all. Carbon dioxide, which is more easily liquefied than most gases, deviates very considerably from Boyle's law, while hydrogen and helium, under ordinary conditions, behave very nearly as ideal gases.

In the algebraic expression of Boyle's law, $pv = b$, the small v indicates the volume per unit mass, or specific volume, while the capital letter V is usually reserved to indicate the total volume. Since density is mass per unit volume, it is the reciprocal of v , and Boyle's law could have been written $p/d = b$, or $p = bd$, indicating that the pressure varies directly as the density, at constant temperature. Boyle's law may also be expressed by $p_1v_1 = p_2v_2$, where the subscripts refer to the pressures and specific volumes before and after a change of pressure.

Since at constant temperature the volume of an ideal gas decreases uniformly with uniformly increasing pressure, it acts like a perfectly elastic helical spring under compression. In fact, a cylinder and tightly fitting piston resting upon an air cushion, would act like the spring, if there were no leakage or friction. In the case of the gas however, the *strain* is due to a volume change, while the spring

changes shape only when it is compressed. The graph of the behavior of such bodies is a curve known as a *rectangular hyperbola* which is asymptotic to the XY axes. Thus Boyle's law is shown in Fig. 101 by the solid line which is plotted from $p v = b$, on the pressure-volume diagram. The dotted line is like the straight portion of the curve in Fig. 84. It represents the stress-strain relations of an ideal gas, when distances along the X axis measure strain, $\Delta V/V$, instead of specific volume. The fact that this line is straight means that ideal gases have perfect bulk elasticity.

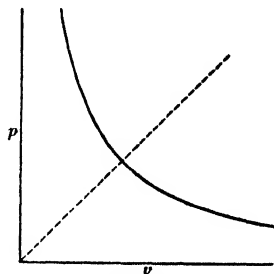


Fig. 101.

147. Work of compression, expansion, and displacement. If the perfectly fitted piston assumed in the last paragraph is forced in against the pressure p , it does work on the gas and stores up potential energy there similar to that in a compressed spring. When the temperature is constant the process is said to be **isothermal**, and the work done may be obtained by the use of the calculus, giving $W = b \log_e (V_1/V_2)$, where V_1 and V_2 are the initial and final volumes, and \log_e means the Napierian or natural logarithm. As Napierian logarithms are approximately 2.3 times as large as ordinary (Briggs) logarithms, the expression for work done may be written

$$W = 2.3 \, p v \log_{10}(V_1/V_2), \quad (1)$$

where p and v are the pressure and corresponding specific volume at any point on the curve. If p is in dynes/cm², and v is in cm³/g, W is in ergs per gram. This work stores up potential energy in the gas and may be recovered by allowing it to expand back to the original volume at constant temperature.

When there is no expansion, and gas or other fluid is supplied at a constant pressure p as the piston moves outward through a distance l , the work done is

$$Fl = pAl = p\Delta V, \text{ or } W = p(V_2 - V_1).$$

If additional gas is not supplied as the piston moves outward, the pressure may still be kept constant by raising the temperature. Then the gas expands and does work equal to $p\Delta V$, as before.

If a compressed gas expands *slowly* into a region of lower but constant pressure, as into the free atmosphere, the process is approxi-

mately isothermal. Then by Boyle's law, $p_1v_1 = p_2v_2$. Subtracting p_2v_1 from each side, we have

$$p_1v_1 - p_2v_1 = p_2v_2 - p_2v_1,$$

whence

$$v_1(p_1 - p_2) = p_2(v_2 - v_1),$$

or

$$v_1\Delta p = p_2\Delta v. \quad (2)$$

But since $p_2\Delta v$ measures the work per unit mass done by the expanding gas under the assumed conditions, $v_1\Delta p$ must measure this work also. In other words, when a gas expands isothermally into a region of constant pressure, the work per unit mass done in pushing aside the surrounding medium is measured by the product of the original specific volume and the difference between the initial and final pressures

$$\text{or} \quad W = v_1\Delta p. \quad (3)$$

As liquids do not expand like gases, they can do work on a piston only when additional liquid is supplied as the piston moves outward. If it is supplied at constant pressure, the work done is $p\Delta V$, as with gases. When unit volume of an incompressible liquid is introduced into the cylinder, $\Delta V = 1$, and the work done is numerically equal to the pressure. Thus the pressure is a measure of the work done upon the liquid wherever it displaces a unit volume against a constant opposing resistance.

A similar situation arises in filling a tank with a liquid by means of a pipe entering from below. Here the opposing resistance is caused by gravity, and the work required to introduce each new unit of volume is numerically equal to the pressure hdg , where h is the height of the liquid column. But as d is the mass of unit volume, hdg is the gravitational potential energy of unit volume of the liquid at the top of the column. Thus the work done in introducing a unit volume appears as increased potential energy.

148. Manometers. In order to measure the pressure of gases, various types of gauges are in use, the simplest forms being called **manometers**. The open-tube manometer is shown in Fig. 102. It is a U-tube open at one end, and with the other connected by flexible tubing to the container of the gas whose pressure is to be measured. In (a) we see it when the gas pressure is the same as that of the atmosphere denoted by P . In (b), the gas pressure is less than P , and in (c) it is greater. If the liquid is of known density, the actual pressure is easily measured. Thus, in case (b), if the liquid is

mercury, the recorded pressure in "millimeters" is the barometric height H at the time of reading, less the difference of level $2h$, both expressed in millimeters. This may be reduced to atmospheres

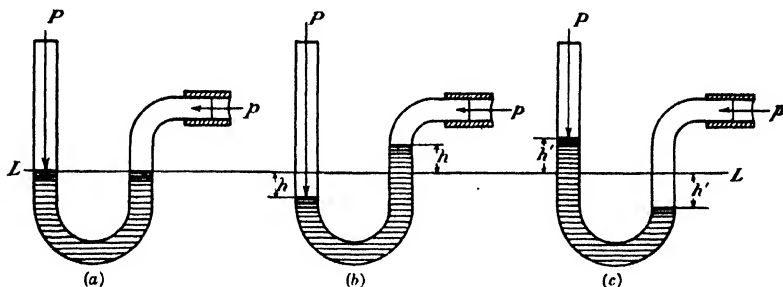


Fig. 102.

by $p = (1 - 2h/H)(H/760)$. If the gas has zero pressure (a perfect vacuum), $2h = H$, and the preceding expression reduces to zero. The calculation may be simplified by expressing the pressure in bars. Then the observed pressure is given by $p = P - 2hdg$.

149. Diving bell and caisson. A diving bell resting on the bottom of a body of water of depth D is partly filled to a height h with water which rises against the increasing pressure of the air in the bell, as indicated in Fig. 103. The internal pressure is found by applying Boyle's law. Thus if the bell's height is H , its section A , and the original atmospheric pressure within it P , we have $PAH = PV$. This must equal the new pressure P' , times the new volume, or $PV = P'A(H - h)$. Equating these values and solving for P' , we obtain

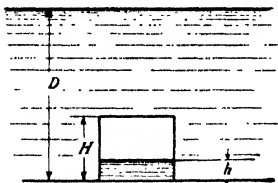


Fig. 103.

$$P' = \frac{PH}{H - h}. \quad (1)$$

The pressure of the water at the depth D is given in bars by Ddg , and at the level of the water in the bell it is $(D - h)dg$.

Then the total pressure acting on the gas is the sum of the atmospheric pressure and the pressure of the liquid at the depth $D - h$ giving

$$P' = P + (D - h)dg. \quad (2)$$

Equating this with the value of P' obtained above, we have

$$P + (D - h)dg = \frac{PH}{H - h}.$$

Solving for D and simplifying gives

$$D = \frac{Ph}{(H - h)dg} + h, \dagger$$

by which the depth may be found from data available within the diving bell if P is known.

This principle is used in a certain type of sounding lead consisting of a hollow tube coated internally with a soluble pigment. When drawn to the surface the distance h to which the water is seen to have entered gives the depth to which it descended, regardless of the inclination of the tube and line.

Caissons used in the construction of bridge foundations are essentially diving bells, but the column of water which tends to enter them is kept out by an increased pressure continuously maintained by compression pumps operated on land. These pump air through pipes down to the caisson. The added pressure required is only that represented by the height h , so that the men working in the caisson breathe air at a pressure corresponding to the total depth D . This is one atmosphere at the surface, two at a depth of 33.9 ft., three at 67.8 ft., and so on at still greater depths.

150. The air pump. There are many devices for exhausting gas from a container and producing a more or less ideal vacuum, though even the best pumps are incapable of producing an absolute void. With a few exceptions like aspirators, mercury vapor pumps, and the molecular pump, all air pumps may be said to operate on the principle of creating a space into which rushes the gas to be exhausted. It is then cut off from its source, and is later allowed to escape into the atmosphere. Toepler's pump may be taken as typical of all of this kind and, though little used today, it is here described because its simplicity makes it easy to understand the basic principles involved. It is shown in its simplest form in Fig. 104. The glass tube B connects the container to be exhausted with the tube J and the bulb M , the latter being drawn out into the narrow-bore barometric tube C which dips into the mercury-filled cup D . A cup G filled with mercury is connected to J by a rubber tube T , so that with atmospheric pressure in M , the mercury in G and J stands at the same level. When G is raised, the mercury in J rises into M which it fills, thus forcing the air trapped there down through C , whence

† When the pressure is measured in gravitational units, P must be multiplied by k to reduce it to dynes/cm² or poundals/ft². This results, numerically, in eliminating g from the denominator. (See footnote, Article 131.)

it bubbles out through the mercury cup. A valve at V prevents the main supply of mercury from rising through B past that point into the gas container, but it allows free access of the gas into M , when the latter is empty. The cup G is now lowered, the mercury level falls in M creating a partial vacuum there, so that the column in C rises to a height h , and the levels of the mercury in the pump proper now differ by the same amount. But when the mercury in the central tube falls below the inlet of B , a rush of gas from the container lowers its pressure and raises somewhat the pressure in M , though not of course back to its original value. This procedure may be repeated as often as desired, until the pressure has fallen to the required value.

The basic principle of an air pump is readily demonstrated as follows: Let us ignore the volumes of the tubes and consider only those of the bulb M and a container N (not illustrated) which is to be exhausted. Let V and V' be the respective volumes of M and N ; then the total volume of gas to be exhausted is $V + V'$. After the first upstroke the total volume is only that of the container (or V'),

which is filled with gas still at atmospheric pressure. After the downstroke, however, this gas expands to fill $V + V'$ once more, and its pressure, from Boyle's law, falls to $p_1 = p_0 V' / (V + V')$, where p_0 is the original atmospheric pressure. The next upstroke again reduces the total volume to V' , but after the following downstroke, it expands to $V + V'$ as before. Then the new pressure $p_2 = p_1 V' / (V + V')$. Substituting for p_1 its value obtained above, $p_2 = p_0 (V')^2 / (V + V')^2$, and after n strokes,

$$p_n = p_0 \left(\frac{V'}{V + V'} \right)^n$$

As a numerical illustration, suppose the volume of the bulb M to be equal to that of the container N . Then $V' / (V + V') = \frac{1}{2}$, and

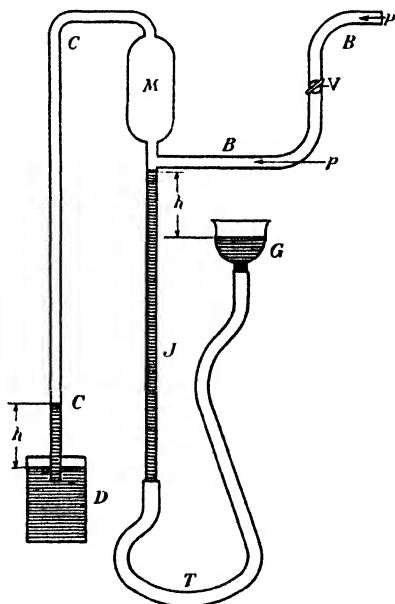


Fig. 104.

after six strokes $p_6 = p_0/64$, or about 12 mm of mercury, if p_0 were normal atmosphere. This is a very poor vacuum, but after 20 strokes, for instance, the pressure should fall to less than a millionth of an atmosphere. The actual result, however, is far short of this ideal for many reasons, though the calculation gives at least some idea of the process of evacuation in the more usual types of air pump.

151. Buoyancy of gases. Archimedes' principle applies to bodies immersed in a gas just as much as to those in a liquid, and all bodies in our atmosphere weigh a little less than they would in a vacuum. This may be demonstrated by balancing a hollow brass ball against an ordinary brass weight, and then placing the balance under the receiver of an air pump (Fig. 105). As the air is exhausted, the pan containing the weight goes up, and the ball goes down, because the latter really weighed more when the balance was adjusted. But it displaces so much

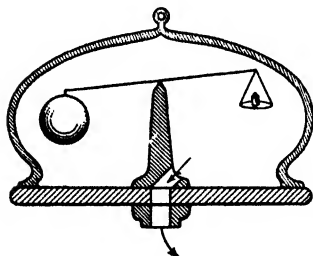


Fig. 105.

more air than the weight, that it is more buoyed up, and appears to weigh the same until the surrounding air is removed.

The preceding experiment shows that objects when balanced against brass weights appear to weigh less or more than they should according to whether they are less or more dense than brass. This error may be allowed for by applying Archimedes' principle. We obtain the true weight from $w = w'd'/(d' - d)$, where w' is the rated value of the brass weights in air, and d' and d are the densities of the object and of air respectively.

SUPPLEMENTARY READING

C. W. C. Kaye, *High Vacua*, Longmans, Green, 1927.

PROBLEMS

1. An ideal gas under atmospheric pressure occupies 5 l. If the pressure is reduced to 20 cm of mercury without change of temperature, what is the volume of the gas? *Ans.* 19 l.

2. A cylinder whose diameter is 8 in. contains an ideal gas at a pressure of 18 lb./in². when the piston is 16 in. from the closed end. What is the pressure and how much work is done after it has been slowly pushed in 10 in., assuming constant temperature? *Ans.* 48 lb./in².; 1181 ft.-lb.

3. The air in a liter container is under a pressure of 6 atmospheres. If its normal density is 0.0012 g/cm^3 , what is the mass in the container? What is its potential energy above that of the outer air at the same temperature? *Ans.* 7.2 g; 1088 joules.

* 4. If the pressure of the air in Problem 3, as the result of a slow leak, falls to two atmospheres, how much gas has escaped? How much energy has been lost? *Ans.* 4.8 g; 948 joules.

5. In an open manometer as shown in Fig. 102 (b) the difference of level of the mercury columns is 24 cm. The barometer reads 75 cm. What fraction of a normal atmosphere is indicated? *Ans.* 0.671.

6. An open manometer as in Fig. 102 (c) contains water. The difference of level when it is connected to a fuel gas main is 6 in. Calculate the total gas pressure in pounds per sq. in. if the barometer reads 30.56 in. of mercury. *Ans.* 15.26 lb./in².

7. A diving bell whose inside height is 7 ft. rests on the bottom of a lake of such depth that the water rises to a height of 4 ft. from its lower rim when the barometer reads 29 in. What is the depth of the lake? *Ans.* 47 ft. 8 in.

8. A block of cork whose density is 0.22 g/cm^3 weighs 286 g in air as shown by brass weights (correct in air) whose density is 8.4 g/cm^3 . Calculate the absolute weight in vacuo, taking air density as 0.0012 g/cm^3 . *Ans.* 287.57 g.

9. The volume of a balloon is 500 m^3 . It is filled with hydrogen whose density is 0.089 g/l . The density of the surrounding air is 1.250 g/l . What is the total lifting power of the gas? *Ans.* 580.5 kg.

CHAPTER 11

Fluids in Motion

152. Rate of flow. The "delivery" of a pipe, or rate of flow through any other channel for fluids, is usually measured in terms of the volume which passes a certain fixed cross section of the channel during some unit of time, as gallons per minute, cubic centimeters or liters per second, and so forth. If the velocity of the fluid at the section s in Fig. 106 is u , then the distance l through which the fluid stream moves in the time t , is ut . This may be regarded as the length of an imaginary cylinder which has passed the section s in the stated time. Taking a as the area of the cylinder section, then its volume $al = aut$, and the volume rate of flow is given by

$$R = aut/t = au.$$

If the rate R' in units of mass per second is desired, then

$$R' = aud,$$

where d is the density of the fluid at s . These values of volume or mass rate of flow, au and aud , can always be used whenever u can be regarded as substantially uniform across the section s , but unfortunately, friction between conductor and fluid, and viscosity within the fluid, result in a variable velocity which is maximum along the axis and decreases toward the bounding surface. So the uniformity of u across s is only approximately realized for large sections, and in general an average velocity has to be used in such calculations.

153. Flow of liquids. The familiar saying, "still waters run deep," expresses an important truth regarding flowing liquids. The deep channel means a greater section than a shallow channel of the same width. Consequently a given number of gallons or pounds per minute move at a slower rate where the stream is deeper. This follows from the incompressible nature of liquids. So in a stream which is flowing through a conductor of varying section, the volume or mass rate of flow must everywhere be the same. If a_1 and a_2 are

two section areas, then as the volume rates are the same, $a_1u_1 = a_2u_2$, or the velocities vary inversely as the section areas.

154. "Head" and pressure gradient. Liquids flow through conductors only because something pushes them. The difference of level between the free surface of water in a tank and a delivery outlet illustrates one possible source of the *push* required. It is called the "head" and is measured in feet or meters. Obviously water under a head of many feet tends to move faster than when its head is only a few inches. It is, however, misleading to express the cause of liquid flow in units of length. A meter cannot move a column of water. It is really the pressure, caused by head, which is acting, and head is merely a convenient measure of gravitational potential which causes the pressure, as was shown in Article 66.

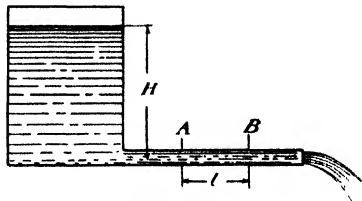


Fig. 107.

If a liquid is moving through a horizontal pipe under a total head H , as in Fig. 107, the friction it encounters causes a steady loss of pressure which may be regarded as a loss of head. This results in a decreased rate of flow. The progressive decrease of pressure along the tube is called the pressure gradient, and is measured by the pressure difference between two sections as A and B divided by the distance l between them. It is the *space rate of change* of pressure.

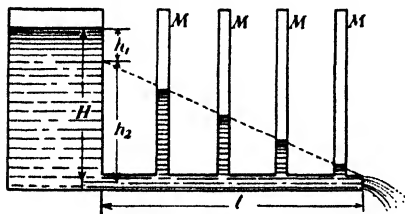


Fig. 108.

The total fall of pressure through the pipe, caused by friction or viscosity, is known as the "friction head." The actual head must exceed the friction head to establish a steady flow, and this necessary excess which actually moves the liquid is known as the

"velocity head." These terms may be illustrated as in Fig. 108, where the height of the liquid in the manometers M indicates the pressures along the horizontal pipe. The total head H is made up of two parts: h_2 , which overcomes the viscosity of the flow and is therefore the friction head, and h_1 , which is the excess required to keep the liquid in motion (or the velocity head), while the pressure gradient is h_2/l , where l is the length of the pipe.

155. Bernoulli's theorem. Let us consider a unit volume of an incompressible liquid, moving without viscosity, with velocity u_1 as it passes the imaginary boundary D in the tube shown in Fig. 109. When it moves so as to displace its own volume, the work done upon it by the liquid pushing it from behind, as explained in Article 147, is numerically equal to p_1 , the pressure indicated by the manometer.

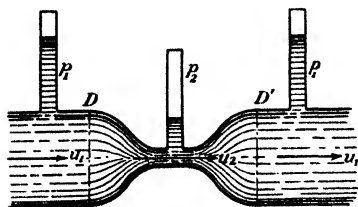


Fig. 109.

We shall further assume that the flow is horizontal, so that no work is done by or against gravity.

Since the liquid in the tube is supposed incompressible, the displacement of unit volume at D involves a displacement of the same volume in the narrow, or *constricted* portion of the tube. Here we must consider the work done by the displaced liquid upon the liquid in front of it, and as usual, this work is measured by the pressure p_2 in the constriction. Thus the net amount of work done in displacing unit volume is the difference between the work done in the two regions we are considering, $p_1 - p_2$. This work appears as a gain in the kinetic energy of the liquid, for it is obvious that it is moving with an increased velocity in the constricted portion of the tube. Using u_1 and u_2 to indicate these two velocities, as shown in Fig. 109, we find that the kinetic energies of the unit volumes are $du_1^2/2$ and $du_2^2/2$ respectively. Then equating the work done by the pressure difference to the gain in kinetic energy, we obtain

$$p_1 - p_2 = du_2^2/2 - du_1^2/2,$$

$$\text{whence} \quad p_1 + du_1^2/2 = p_2 + du_2^2/2, \quad (1)$$

$$\text{or in general,} \quad p + du^2/2 = \text{constant}. \quad (2)$$

This is Bernoulli's theorem for horizontal flow. It applies equally to the liquid after emerging from the constriction into the wider section at D' . But in this case, the second pressure is greater than the first, and more work is done in opposing unit volume at D' than is done upon it in the constriction by the liquid pushing from behind. This is because of decreasing velocity and lower instead of higher kinetic energy.

If there is a difference of level h between the two portions of the tube, a third term, hdg , must be added to equation (2) to take account

of the change of gravitational potential energy. Then the general theorem becomes

$$p + hdg + du^2/2 = \text{constant.} \quad (3)$$

Bernoulli's theorem tells us that pressures are least where velocities are greatest, and vice versa. A familiar illustration is the congestion of a crowd before a narrow passage, where there is almost no motion. But this is followed by greatly increased separation of the individuals, with increased speed, as they go through the passage.

In liquids the conclusion just arrived at might have been predicted from the fact that the higher speed which is inevitable in the constriction involves acceleration, and this can be produced only if the pressure behind is greater than that in front. The reverse is true when the tube widens out again, for then the acceleration is negative, and the pressure in front must be greater than that behind.

156. Aspirators. Perhaps the most valuable application of Bernoulli's principle is the steam injector, a type of condenser associated with steam engines. As shown in Fig. 110, a jet of steam issuing from the nozzle is at a reduced pressure, and the exhaust steam from the cylinder tends to rush into this space through the side tube *T*. There it is entrained by the jet and ultimately ejected at *K*. The aspirators used in laboratories to produce a partial vacuum operate in much the same way; but the jet is of water, and it carries off air from the side tube instead of condensed steam.

The general principle that fluids moving at high speeds experience reduced pressure has many other applications and can be demonstrated in a variety of ways. The atomizer used in spraying perfumes operates by causing a strong air jet to lower the pressure over a tube which dips into the perfume.

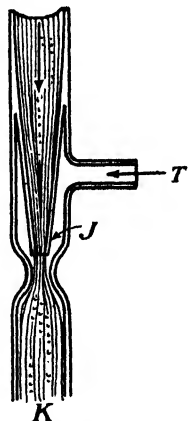


Fig. 110.

The curved path of a spinning baseball may also be explained in a similar manner. We may imagine the ball to be at rest in a strong air current which streams past it. This represents a ball pitched without a spin where, if gravity is neglected, the only force acting is due to the relative motion of the ball and the air. If it is then set spinning, it drags a layer of air around with it because of viscosity. This is shown in Fig. 111, where the arrows *AA* represent the relative velocity of the air streaming past the ball. The curved arrows *aa'* represent the velocity of a layer of air set whirling by the ball's spin. The resultant of the velocities *A* and *a* is *R*, which is less than *A* be-

cause on this side the air close to the ball is moving in opposition to the main stream. Similarly R' is greater than A , because on this side A and a' help each other. The result is an increased pressure on the side where the resultant velocity is least and a decreased pressure where it is greatest, with a force acting from R toward R' . This gives the ball an acceleration in the direction of the vector F , and causes it to move in a curved path as indicated by the arrow V .

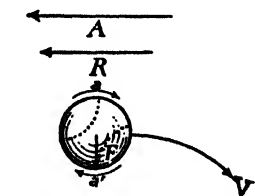


Fig. 111.

A light celluloid ball may be supported in an upward air blast from a small orifice, because within the blast the pressure is reduced below that of the atmosphere outside, and any tendency to get outside the jet is counteracted by the higher pressure there which forces it back.

In the same way a vertical jet of water retains a ball of suitable weight and size apparently balanced upon it.

When two boats are anchored near each other in a swiftly flowing river they tend to swing together. This is caused by the necessarily increased rate of flow in the narrow space between the boats, thus reducing the pressure there, so that they are forced together by the greater pressure outside. The same is true if the boats are steaming along side by side through still water, because it is only the relative motion which counts.

An interesting experiment which illustrates these phenomena is easily performed by blowing upward through the lower end of a spool (held vertically) against a stiff card which rests upon the upper end. If the card is prevented from slipping sideways by a pin stuck through its center and going down into the hole in the spool, the harder one blows, the more the card hugs the surface.

157. Torricelli's theorem. If a vessel containing a liquid has an orifice below the surface, the liquid rushes through it with a certain velocity u , and the jet describes the arc of a parabola like any other projectile, as shown in Fig. 112. This velocity may be calculated by equating the kinetic energy acquired by unit volume of the liquid with the potential energy lost by the system. The mass of a unit volume is the density d ; therefore the kinetic energy of this mass as it issues from the outlet is $du^2/2$. The lost potential energy is equal to the work required to introduce the

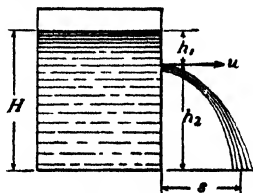


Fig. 112.

unit volume at the level of the outlet, at a distance h_1 below the surface. It therefore equals the pressure $h_1 dg$ at that level, as explained in Article 147. Equating these energies, we obtain

$$\begin{aligned} du^2/2 &= h_1 dg. \\ \therefore u &= \sqrt{2gh_1}, \end{aligned} \quad (1)$$

which is seen to be the same as the vertical velocity that all bodies acquire after falling freely through the same distance. Equation (1) may be transformed by multiplying and dividing by the density. Then

$$u = \sqrt{\frac{2gh_1 d}{d}} = \sqrt{\frac{2p}{d}}, \quad (2)$$

where p is the pressure. Thus the velocity of the jet at any level may be found from the pressure at that level.

Torricelli's theorem would seem to give a measure of the delivery of such a jet if the area A of the orifice were known, for then the rate of flow R , in units of volume per second, would be Au . But this is not the case, because the lines of flow within the vessel result in a contraction of the jet ("vena contracta") outside, so that the effective area of the orifice is a instead of A , as shown in Fig. 113. This difficulty, however, can be considerably obviated by using a specially shaped nozzle instead of a hole in the side of the container.

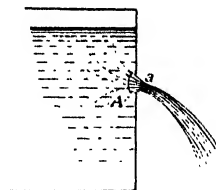


Fig. 113.

The range s of the jet in Fig. 112, issuing horizontally at a height h_2 above the plane, is found by calculating the time it takes to fall this distance. The time is obtained from $h_2 = \frac{1}{2}gt^2$, giving $t = \sqrt{2h_2/g}$. By Torricelli's theorem, $u = \sqrt{2gh_1}$. Then substituting these values in $s = ut$, we obtain

$$s = \sqrt{2gh_1} \sqrt{2h_2/g} = 2\sqrt{h_1 h_2}. \quad (3)$$

This range is easily shown to be a maximum when $h_1 = h_2$.

158. Effusion of gases. The case for gases corresponding to Torricelli's theorem is demonstrated when a gas under a pressure p flows, or *effuses*, through a small orifice of section area A out of a container into the atmosphere. The kinetic energy of a mass m of the jet is $mu^2/2$. Let m represent the mass that issues from the orifice in the time t . Then its volume V is found by imagining a cylinder of the gas, whose length is ut , passing through the hole whose section area is A . The volume of this cylinder is Aut , and $m = Autd$. So the kinetic energy is given by

$$W_k = Au^3 td/2. \quad (1)$$

The potential energy lost in the same time by the emission of the volume V of the gas may be calculated if the effusion is slow enough to be essentially isothermal. Then in accordance with equation (3) of Article 147,

$$W_p = V\Delta p,$$

or

$$W_p = Aut\Delta p. \quad (2)$$

But the gain of kinetic energy equals the loss of potential. Therefore, equating (1) and (2), we have

$$Au^3td/2 = Aut\Delta p.$$

$$\therefore u = \sqrt{\frac{2\Delta p}{d}}. \quad (3)$$

This equation serves as a valuable means for determining the density of a gas. The difference Δp between the pressures is readily measured, and the velocity u is found from the rate of effusion R , for under ideal conditions

$$R = \frac{V}{t} = \frac{Aut}{t} = Au. \quad (4)$$

Then if a correction term is used to allow for the effective area of the outlet, we may calculate u and so obtain the density.

SUPPLEMENTARY READING

King and Wisler, *Hydraulics* (Chapters 6 and 7), Wiley, 1927.

PROBLEMS

1. A vertical hollow cylinder whose diameter is 8 cm contains water whose level is maintained at 46 cm while it flows horizontally through an orifice at the bottom having an effective area of 1.48 cm². Calculate the velocity of efflux and the average velocity in the tank. *Ans.* 3 m/sec.; 8.84 cm/sec.

2. In Fig. 109, the level of water indicating p_1 is 12 in., and the section at D is 3 in². The section of the constriction is 0.5 in². The rate of flow is 36 in.³/sec., and $d = 0.0361$ lb./in³. Find p_2 . *Ans.* 0.196 lb./in².

3. The cross-section of a siphon is 0.4 cm². It is desired to produce a flow of water from a tank at the rate of 8 l per minute. What is the required distance between the water level in the tank and the lower end of the siphon? *Ans.* 56.67 cm.

4. Gas under a constant pressure of 1.1 atmosphere is forced through a small orifice into the outer atmosphere at the rate of 16 cm³ per sec. The effective area of the orifice is 0.002 cm². What is the density of the gas? *Ans.* 0.0032 g/cm³.

CHAPTER 12

Surface Tension and Capillarity

159. Molecular forces. The mutual attraction between molecules constitutes, as we have seen, the force of cohesion. The cause of this force and its laws are not clear, but it is safe to say that it varies inversely with some higher power of the distance than ordinary gravitational forces and the inverse square law. Consequently it is very great when the molecules are close together, but falls off with extreme rapidity as they separate. This is especially noticeable in the case of solids, for the cohesive forces are the source of their rigidity, and depend upon unbroken continuity of molecules packed closely together. If a bar of iron is sawed in two it cannot be made to cohere again by pressing the ends together, since owing to the roughness of the surfaces in contact, only a very small number of molecules are close enough together to attract each other appreciably. However, these surfaces may be so accurately planed and polished that the deviations from a perfect plane do not exceed a few millionths of a centimeter. Then when pressed together they do cohere with considerable force. Secondary standards of length made of "optically plane" slabs of steel cohere in this way with a force intensity sometimes as high as 30 kilograms per square centimeter, which is about thirty times atmospheric pressure.

160. Molecular range. The greatest distance through which molecules exert a measurable attraction for each other is known as the **molecular range**. It was first measured by Quincke between glass and water (in this case *adhesion* rather than cohesion), and quite recently Chamberlain has determined it with considerable precision. Though molecular range varies with different substances and conditions, it may be taken as of the order of 1.5×10^{-7} cm. Consequently it is safe to say that no molecules attract each other with appreciable force if they are much farther apart than this distance. We may then think of a molecule, whether of a solid or liquid, as surrounded by a sphere whose radius is its molecular range. Within this sphere it attracts other molecules and is attracted by them, while molecules lying outside the sphere have no appreciable influence upon it. In gases under

ordinary conditions the molecules are too far apart to attract each other enough to make their molecular range of any real significance.

161. Surface films. The free surface of a liquid is subject to a certain unbalanced stress which is not shared by the rest. The three circles in Fig. 114 represent the spheres of attraction of the three molecules at their centers. The sphere *A* lies wholly below the surface, and the molecule at its center is attracted with balanced radial forces in all directions. The sphere *B* around another molecule nearer the surface lies partly outside the liquid.



Fig. 114.

its central molecule are also balanced within the central zone, being equal and opposite in pairs, but the heavily shaded zone

at the bottom has no corresponding portion to balance it, except for the very feeble attractions of the air molecules within the unshaded region lying outside the liquid. Finally, *C* is the range around a molecule on the surface, with all the molecules in its lower hemisphere exerting an almost wholly unbalanced pull whose resultant is directed vertically downward. Thus, if we construct a plane indicated by *L* whose distance below the surface is the molecular range r , the molecules above it, lying nearer and nearer the surface, experience an increasing downward pull. The space between these two levels constitutes what is known as a surface film.

All the molecules within the film just defined are under a vertical pull which begins at *L* and is a maximum on the surface, and they therefore possess potential energy due to their favored position. Evidently work must be done on any molecule to move it from below *L* into the region of stress, against an ever-increasing force tending to pull it back, and this work measures the potential energy acquired. But the potential energy of any system tends to become as small as possible; therefore, since the thickness of the film is constant, the number of molecules under stress can diminish only by decreasing the area of the film. Thus the free surface of a liquid acts like a stretched membrane under a constant surface tension T which is measured in terms of force per unit length. In this respect, it is unlike a stretched drumhead, for there the tension increases with the amount of stretch, like any elastic body. In surface films the tension is constant, and has no reference to the total area.

162. Evidence of surface tension. Like a membrane under tension, the surface of a liquid in a jar tends to remain perfectly plane, for then its area bounded by the containing vessel is a minimum. Let a light object which is not wet by the liquid, like a slightly greasy needle, be carefully laid on its surface, as shown in section in Fig. 115 (a). The needle forms a depression in the film whose increased area supplies an upward force. This counterbalances the downward pull of gravity, as indicated by the arrows, and the object floats.

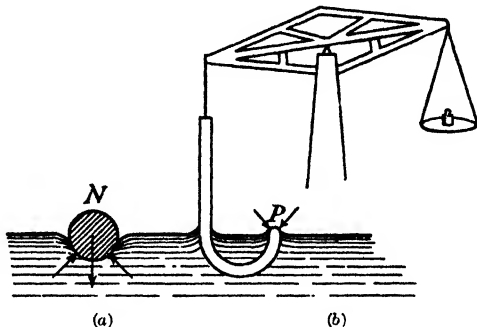


Fig. 115.

The reverse phenomenon is shown in the same diagram (b) where a hook P is accurately balanced with its point below the surface. It is then brought up so that the point tends to break through the film. Until it does so, the minute elevation in the film exerts a downward force on the point, and the equilibrium of the balance is destroyed.

163. Angle of contact. At the edge of a surface film, where it comes in contact with the containing vessel, a new condition develops, due to the adhesion between the molecules of the solid and the liquid. This force differs between different substances, but is notably strong between glass and water, and very weak between glass and mercury, to select extreme cases. In Fig.

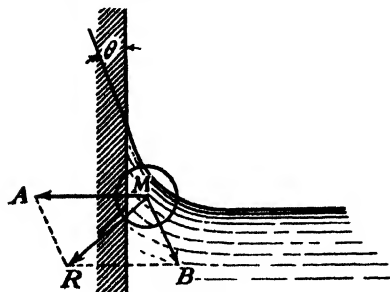


Fig. 116.

Fig. 116 the molecule M , lying within the surface film, is surrounded by the sphere of molecular range which lies partly in the liquid, partly in the solid, and partly in air. Let B represent the resultant pull of the liquid molecules, and A that due to the solid; then neglecting the extremely small effect of the air, we

obtain a resultant R from the force parallelogram. This replaces the vertical pull in the previous illustrations. Since R is much greater than gravity for the molecule M , we may ignore its weight, and construct the film surface perpendicular to R , because free surfaces are

always normal to the resultant force which acts upon them. Thus the liquid surface curves upward in what is known as a **concave meniscus**, and meets the surface of the solid at some angle θ known as the **angle of contact**. This angle is about 8° for clean glass and water, but has other values with other substances.

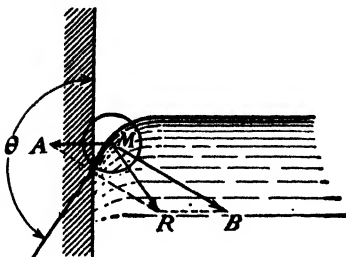


Fig. 117.

In the case just described, the resultant adhesive force A was taken as greater than B , but if it is small compared to B , the situation shown in Fig. 117 arises. Here the resultant R causes the surface of the liquid to curve downward. This is the case with mercury against glass, which is then said to have a **convex meniscus**.

In the case of water against silver, the adhesive and cohesive forces are about equal. Then A equals B , R is vertical, the angle of contact is 90° , and there is no meniscus.

164. Capillarity. If the containing vessel is a tube of small bore, the meniscus may meet the axis so that the whole surface is curved either concave or convex upward. Let us examine the conditions for equilibrium when such a tube, known as a capillary (hair tube), has one end immersed in a dish of liquid. If the liquid "wets" the tube there is a concave meniscus. The surface tension along the bounding circle CC in Fig. 118 is directed outward at an angle θ with the vertical. Unlike pressure, surface tension is measured in terms of *force per unit length*, instead of force per unit area of the film. Therefore, since the tension T acts along the entire circumference $2\pi r$, the total force is $2\pi rT$, and its vertical component is $2\pi rT \cos \theta$. As a result of this upward force the liquid rises in the tube until the downward pull of gravity on the column balances the force due to the tension. The weight of the column is given by $\pi r^2 h d g$, where h is its average height above the level of the liquid outside, and r is the radius of its section. Therefore

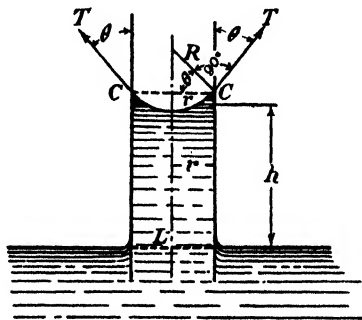


Fig. 118.

$$2\pi rT \cos \theta = \pi r^2 h d g,$$

and
$$h = \frac{2T \cos \theta}{rdg} = \frac{4T \cos \theta}{Ddg},$$

where D is the tube's diameter.

In the case of water against clean glass, $\theta = 8^\circ$, $\cos \theta = 1$ approximately, and

$$h = \frac{4T}{Ddg}.$$

A similar formula applies to the *depression* of a convex meniscus when $\theta > 90^\circ$, and its cosine is therefore negative, as illustrated in Fig. 119.

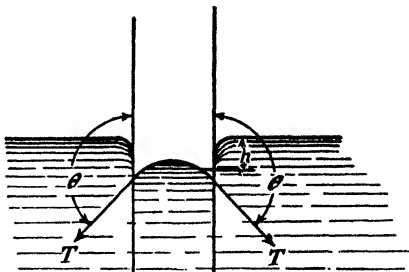


Fig. 119.

In case the bounding surfaces of the capillary column are two planes very close together, like two parallel sheets of glass immersed in a jar of water, then the upward force per unit length measured along the surface of the capillary column is simply $T \cos \theta$ on each side. The weight of such a column per unit of horizontal length is hsg , where s is the distance between the sheets, so that

$$2T \cos \theta = hsg, \text{ and } h = \frac{2T \cos \theta}{sdg},$$

which is half the height in a tube of diameter s .

165. Curvature and pressure. Since the surface of the film in capillaries of small bore is nearly spherical, the radius of this sphere

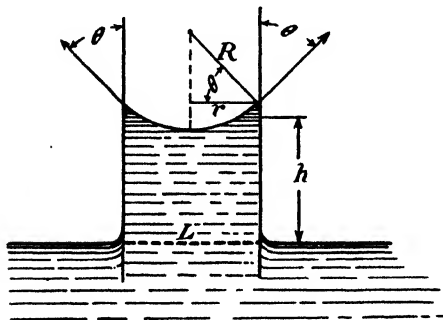


Fig. 120.

and that of the bore are connected by the equation $r = R \cos \theta$, as shown in Fig. 120. Substituting this in the general equation for h , we obtain $h = 2T/Rdg$; therefore $hdg = 2T/R$. But hdg is the *pressure* due to the column of height h , and since the pressure at the level L is one atmosphere, it is hdg less than P just under the film,

and atmospheric again just above it. Thus in passing through the film from its convex to its concave side, the pressure increases by an

amount equal to hdg . Denoting this change in pressure by Δp , it follows that

$$\Delta p = hdg = \frac{2T}{R} = 2T\sigma, \quad (1)$$

where σ is the curvature of the surface, or the reciprocal of its radius.

This change in pressure is a general property of all curved surface films, though the above result is valid for single spherical films only.

It is possible to obtain Δp for a cylindrical surface from the fact, already proved, that the column between parallel plates is only half as high as in a tube whose diameter equals the distance between the plates. As the radius R of the film's curvature is unaltered, the pressure difference, Δp , given in equation (1), now becomes $\Delta p = T/R = T\sigma$. This result is to be expected, because a spherical surface may be regarded as composed of two cylindrical surfaces at right angles to each other, so that the double curvature σ of the sphere is twice as effective as the same single curvature of the cylinder.

166. Double films. Films such as those in soap bubbles are really double, with a relatively thick layer of liquid between them, as suggested in Fig. 121. To produce films of this kind with water, soap must be added. This makes them less fragile, in spite of the fact that the surface tension is thereby decreased. Pure water has a high surface tension, but double films of water alone are very fragile and can be formed only with difficulty.



Fig. 121.

If a double film is exposed to atmospheric pressure on both surfaces, it is known as a *free film*, and is then stretched across an area bounded by some contour made of any solid to which the film adheres.

167. Soap bubbles. The pressure within a soap bubble is in excess of the surrounding atmosphere, because of the surface tension which makes it contract, thus compressing the air within it. It is spherical, because the sphere has the smallest area of any solid enclosing the same volume. The increase in pressure passing within the film from the convex to the concave side, is given by the same expression as the one derived for single spherical films multiplied by *two*, because of the *double* film, so that the total internal pressure is

$$P + \Delta p = P + \frac{4T}{R}.$$

As the bubble swells with blowing, Δp steadily decreases as R increases, and the internal pressure approaches P , that of the atmosphere. This explains why small bubbles still on the "pipe" shrink rapidly

when the stem is open, while large ones do not, because the greater internal pressure in the former case forces the air out more rapidly through the stem.

168. Free films. If a double film is stretched across a wire bent to form a rectangular contour open at one side cd , as shown in Fig. 122, but with a light wire $c'd'$ laid across it to determine the missing side, the film tends to contract and pull $c'd'$ along with it. The force with which it contracts is constant for any length l , and its value is $2Ts$. This can be roughly measured by finding the weight mg which just balances it. If we drop small loops of fine thread or hair on a film stretched across any convenient frame, they may be brought into complete contact with the film, as at A in Fig. 123, without breaking it. Then if the film within the loop is punctured with a pin, the loop suddenly assumes the perfectly circular contour shown at B . This is because the circle has the largest area of any contour of a given length, so by pulling the thread into a circle, the unbroken film attains a minimum area.

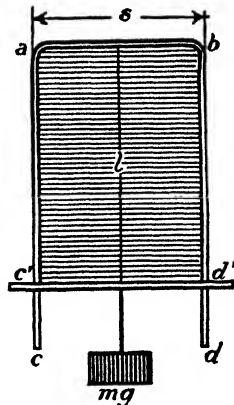


Fig. 122.

169. Minimal surfaces. A free film stretched across any contour, plane or otherwise, has the smallest surface compatible with its boundary. If the contour lies in a plane, the surface is of course plane also, but otherwise the surface is one of double curvature.

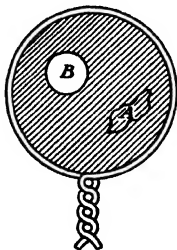


Fig. 123.

This condition can be studied with the aid of the formula $\Delta p = 4T\sigma$, for double spherical films. We may expand the expression to $\Delta p = 2T(\sigma_1 + \sigma_2)$, making it applicable to double curvature, and if the radii are different, $\Delta p = 2T(\sigma_1 + \sigma_2)$, where σ_1 and σ_2 are the curvatures in planes at right angles to each other. Then

$$\Delta p = 2T \left(\frac{1}{R_1} + \frac{1}{R_2} \right).$$

If the film is *free*, the difference in pressure between its two faces is zero, but $2T$ is not zero; therefore

$$\frac{1}{R_1} = -\frac{1}{R_2}.$$

This is the equation of a surface of double curvature, with the two centers of curvature on opposite sides of the surface, and equidistant from it. It is much like a horse's saddle which is concave upward when seen from the side, but concave downward as it curves over the horse's back. Thus in Fig. 124, the cylindrical contour $abcd$ has a radius of curvature R . The film stretched over it sags with two equal and opposite curvatures of radii $R_1 = R_2$, shown at P , and its area is then less than if it stretched straight across to make a cylindrical surface.

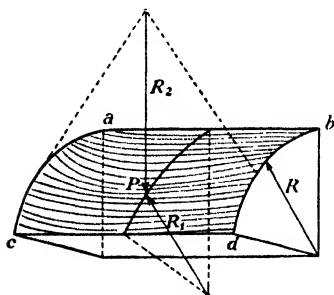


Fig. 124.

170. Films between liquids. If two liquids are in contact, as occurs when a lighter one floats upon a heavier with which it does not mix, the bounding surface film possesses the same characteristic of tension as a film between a liquid and air. But now it is due to the influence of both liquids acting in opposition, and the tension is therefore less than with either liquid alone against air.

If a drop of some lighter liquid c floats upon a heavier one b , as shown in Fig. 125, there are three surface tensions to consider: the ab film between air and the liquid b , whose tension is T_{ab} , the ac film with a tension T_{ac} , and the double liquid film with a tension T_{bc} . These three tensions meet along the circle which bounds the drop at its outer edge, and can be in equilibrium only when their vector sum is zero. This occurs when $T_{ac} \cos \theta + T_{bc} \cos \phi = T_{ab}$. As the tension of the air-liquid film T_{ac} is always greater than the liquid-liquid film T_{bc} , and as their resultant must be horizontal, it follows that ϕ is always greater than θ . The drop is then lenticular (lens-shaped) with the lower surface more curved than the upper.

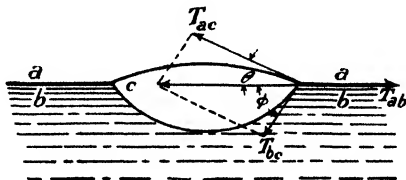


Fig. 125.

If T_{ab} equals or exceeds the arithmetical sum $T_{ac} + T_{bc}$, then θ and ϕ are both zero, and no equilibrium is possible. In the case of petroleum on water, $T_{ab} > T_{ac} + T_{bc}$, and the oil is pulled out by this surface tension as a very thin film over a large area. This property is sometimes made use of during storms at sea. It does not stop

the waves, but by substituting an air-oil film for air-water the much lessened surface tension prevents the formation of a crest with subsequent "breaking" of the wave.

171. Attractions between floating bodies. If two bodies, both wet by the liquid, come near enough to each other so that the liquid rises between them by capillarity, then at some level such as L in Fig. 126, they receive an unbalanced push together. This is due to the fact that the atmospheric pressure P acting on the outer faces is greater than the pressure within the capillary column, because it is above the mean level of the liquid. At other levels, where the opposing pressures are either both inside or both outside of the liquid, they are equal. If the liquid wets neither of the bodies, they are still pushed together, because at the level L (Fig. 127) p is greater

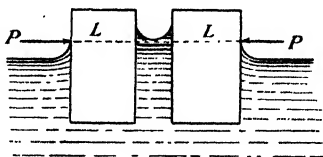


Fig. 126.

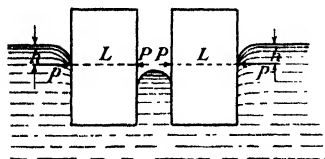


Fig. 127.

than the opposing pressure P , being at a distance h below the mean surface.

Finally, when one body is wet by the liquid and the other is not, the bodies are pushed apart. In this case the surface of the liquid between them has both a concave and a convex

meniscus, as shown in Fig. 128. The concave meniscus on the inside of A does not rise so high as it does on the outside, because it is pulled downward by the capillary depression against the inside of B . The convex meniscus on the inside of B does not go so low as it does on the outside, because it is pulled up by the capillary elevation against the inside of A . Therefore between the levels L_1 and L_2 marking the upper boundaries of the liquid against the two faces of A , there are unequal pressures. The pressure p_1 is less than the atmosphere because it is above the mean level of the liquid. The pressure P opposed to it is that of the atmosphere, and being greater than p_1 , the body A is pushed to the left. Similarly B experiences a push to the right because between the levels l_1 and l_2 the pressure p_2 is below the mean level of the liquid and is therefore greater than the atmospheric pressure P which opposes it.

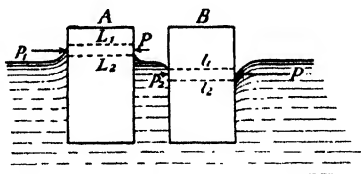


Fig. 128.

Surface Tensions (dynes/cm), at 20° C, and Angles of Contact (liquid-glass)

Liquid	Tension	Angle
Water-air surface	73	8° to 9°
Soap Solution-air surface	30 (approx.)	—
Alcohol-vapor surface	22.3	0°
Ether-vapor surface	17.0	16°
Paraffin Oil-air surface	26.4	26°
Mercury-air surface	465.	130° (approx).
Mercury-water surface	375.	—
Paraffin Oil-water surface	48.3	—
Olive Oil-water surface	18.2	—

SUPPLEMENTARY READING

C. V. Boys, *Soap Bubbles*, Macmillan, 1928.

H. A. Erikson, *Elements of Mechanics*, McGraw-Hill, 1927.

Physics in General

W. F. Magie, *A Source Book in Physics*, McGraw-Hill, 1935.

H. Buckely, *A Short History of Physics*, Van Nostrand, 1927.

Philipp Lenard, *Great Men of Science*, Macmillan, 1934.

PROBLEMS

1. How high does water rise in a capillary glass tube whose diameter is 0.044 cm, if the angle of contact is negligibly small? *Ans.* 6.7 cm.

2. How high would paraffin oil rise in a glass tube whose diameter is 0.036 cm? *Ans.* 3.4 cm.

3. Calculate the depression of a mercury column in a glass tube whose diameter is 0.058 cm. *Ans.* 1.8 cm.

4. What is the pressure due to surface tension inside a soap bubble whose diameter is 6 cm? *Ans.* 40 dynes/cm².

PART II
HEAT

CHAPTER 13

Temperature

172. Hot and cold. Everyone knows the difference between hot and cold water, or a hot and a cold day. But a rigorous definition of such distinctions is by no means easy. We speak of the *temperature* of the air, but have no very clear idea of what kind of measure temperature really is. In fact, it is almost as difficult to define as time, and is therefore best regarded as a fundamental concept which cannot be expressed in terms of the three fundamental “dimensions”—distance, mass, and time. Temperature then is to be regarded as a new unit of measure, the significance of which will become clearer as we study its measurement and its effects.

Like distance, mass, and time, temperature is measured in terms of itself, while the notions of hot and cold are purely relative terms which may be compared by some arbitrary unit such as the degrees of a thermometer, just as two masses are compared by a common unit of measure, the gram.

It is then meaningless to call something merely hot or cold. The boiler of a steam engine filled with the water from a “hot” bath would be considered cold, while a cake of ice is red hot compared to liquid air and causes it to boil violently. So science generally avoids such words as hot and cold, and prefers their comparatives, *hotter* and *colder*. But even then, how can we know with certainty that one thing is hotter than another? A cork mat in a swimming pool *feels* warmer than the porcelain tiles, though really it must be at substantially the same temperature. Evidently our sensations are a very poor guide. Of course we can appeal to the testimony of a thermometer, but what assurance have we that it gives us a fair comparison between the temperatures of different objects?

173. Comparative temperature. In order to answer the preceding questions, we must understand a fundamental fact to be discussed farther on as “the second law of thermodynamics.” It is a matter of universal experience that a group of objects at different temperatures tends to reach one common degree of warmth. The hotter

bodies cool down, and the colder ones warm up. No one doubts this fact, and yet it cannot be said to be a law of nature, any more than the statement that more persons will die in New York City next year than in some country village. Of course we may assume that they will, but it is only a matter of overwhelming probability, for after all, no basic law of nature demands that anyone die within a given year.

So we know from universal experience that hotter bodies automatically warm up cooler bodies, and the probability that this will occur is so inconceivably great that the reverse process may be as absolutely ignored as if it actually were impossible. It is, however, not unthinkable, as will be explained later.

We thus have a sure test of relative hotness and coldness, and may use our instrument of comparison, the thermometer, with confidence, for we know that it will assume the temperature of the object whose temperature is to be determined, provided it is left long enough in contact with it to allow the combined system (thermometer and object) to arrive at a common temperature.

174. Temperature level. From what has been said, we are justified in regarding temperature as a measure of heat level, just as the readings of a barometer at different altitudes may serve to compare levels above the surface of the earth in terms of differences in atmospheric pressure. Or difference of altitude may be defined in terms of the tendency of water to flow from one level to another.

One meadow is higher than another if it drains its surplus moisture into the lower one. So the heat of one object is at a higher temperature level than that of another if it shows a greater tendency to cool off under conditions that are identical, except for the temperature difference.

175. Thermometric properties. Any measurable effect due to heating a body may be used as the basis of a thermometer, and as nearly all the physical properties of matter are altered by heating, there are a great many possible thermometers. Glass, for instance, softens progressively as it is heated, and a measure of its softness might conceivably be used to measure its temperature. But among the numerous possible changes due to changing temperature, only five have been found sufficiently dependable and easy to observe to be suitable for practical purposes. These are changing volume, changing pressure (of gases), changing electrical resistance, changing thermoelectromotive force, and the changing color of bodies hot enough to be luminous. These will all be discussed in their proper places, but for the present we need consider only one—change in volume.

176. Thermal expansion. Most bodies increase in bulk when heated. There are a few important exceptions, but in general, heat results in expansion, as is well known from many everyday phenomena. The draught up a chimney is caused chiefly by the expansion of the heated air, which thus becomes lighter than the surrounding atmosphere and rises. Iron tires are heated before placing them on the wheel, so that they may shrink on and hold tight when cold. Hot-water heating systems are equipped with expansion tanks near the top of the house to provide for the increased volume of the liquid as it is warmed by the furnace.

It is this property which is used in the ordinary thermometer, whether the substance used is mercury, alcohol, or a strip of metal.

177. The mercurial thermometer. Mercury is one of the best liquids for thermometric purposes. It freezes at 38.9 degrees below zero on the centigrade scale, which is a point seldom reached except in the arctic regions. It expands by nearly equal amounts for each equal interval of temperature, until it reaches relatively high levels. It is opaque and readily seen against glass. It is easily purified, and does not stick to the glass tube which contains it. Finally, mercury assumes the temperature of its surroundings relatively quickly, and its expansion per unit rise of temperature is fairly large, though it is surpassed in this respect by some liquids such as glycerine and toluene.

The mercurial thermometer has a bulb blown at the end of a glass tube of small bore. This is exhausted before sealing, so that the space above the mercury is a fairly good vacuum, which prevents oxidation of the surface of the column, and the possibility of breaking the tube by compressed air above the rising mercury.

Since the bulb contains so much more mercury than the stem, almost all the total expansion is due to its contents, and the capillary column above it really acts mainly as an index, more and more liquid being forced into it as the main mass gains in volume and seeks an outlet. Actually the glass bulb expands at the same time as its contents, but not nearly so much, and any error involved in this way is allowed for when the scale is laid off on the stem.

It is evident that such a thermometer may be made extremely sensitive to small changes of temperature by having an especially large bulb, and stem of very small bore, thus magnifying the changes in volume. In this way thermometers are designed to read to hundredths of a degree.

178. Thermometric scales. In order to use the thermometer as a measure of difference of temperature level it is necessary to decide

upon a scale, which involves the choice of the scale division and some arbitrary zero at which the scale begins.

In the centigrade scale, designed by the Swedish astronomer, Anders Celsius, about 1742, the zero is at the freezing point of water, and 100 marks the boiling point under the standard atmospheric pressure of 76 cm of mercury. Thus the degree is defined as one hundredth of the difference of temperature between these two "fixed points." The ease with which centigrade thermometers may have these points checked to test their accuracy, and the fact that temperatures most commonly met with lie within this range of one hundred equal divisions, make this scale the best one for all but a few purposes. It is used by men of science everywhere, and by the people generally in all but the English-speaking nations and a few others.

In the United States and Great Britain the conservative temper of the people has kept in vogue the older Fahrenheit scale devised about 1709 by the German natural philosopher of that name. Its zero was determined by the temperature of a certain cold winter's day in Danzig, where its inventor resided, and was later specified as the temperature of a given mixture of snow, sal ammoniac, and common salt. The other fixed point was obtained from the supposed normal temperature of the human body, and the interval between the two was ultimately divided into 96 degrees. This choice, combined with the arbitrary zero, made freezing at the 32° mark on the scale, and boiling at 212° . In spite of its illogical nature, this scale has the advantage of having smaller degrees than that of Celsius, thus avoiding the use of fractions of degrees unless real precision is needed, and the further advantage that the weather of the temperate and torrid zones is rarely cold enough to bring about negative readings, below zero.

A third scale, invented by the French philosopher Réaumur in 1731, and named after him, is still somewhat used in Scandinavia, Germany, and Holland. Its zero is at the freezing point of water, and the boiling point corresponds to 80° on the scale. Thus its divisions are still larger than centigrade degrees and there is nothing gained by using it in preference to the latter scale.

179. Conversion of temperature scales. Because the less scientific scales of Réaumur and Fahrenheit are still used in some parts of the world, it is often necessary to convert temperature readings between them and the centigrade scale. The conversion between Fahrenheit and centigrade is obtained as follows: Since the interval between freezing and boiling is 180° in one, and 100° in the other, the magni-

tude of the degree is as 180 is to 100, so their ratio is 9:5. Therefore n degrees centigrade equal $9n/5$ Fahrenheit. But as 0° centigrade corresponds to 32° Fahrenheit we must allow for this also; therefore

$$n^\circ \text{C} = \left(\frac{9}{5}n + 32\right)^\circ \text{F}, \text{ or } n^\circ \text{F} = \frac{5}{9}(n - 32)^\circ \text{C}.$$

Thus 10° centigrade $= \frac{9}{5} \times 10 = 18$ Fahrenheit scale divisions, but as the zero of the latter stands 32° below the centigrade zero, we must add 32, making 10° centigrade $= 50^\circ$ Fahrenheit. Or taking -13° Fahrenheit we must add -32° to find the number of Fahrenheit degrees -13° is below freezing. The result, -45° , is then reduced to centigrade degrees, giving $-\frac{5}{9} \times 45^\circ = -25^\circ$ centigrade. The method of converting Réaumur to Fahrenheit and vice versa is the same except that the ratio of degrees is $\frac{80}{180} = \frac{4}{9}$ instead of $\frac{5}{9}$.

It is interesting to note that the centigrade and Fahrenheit scales agree at -40° . This is found by setting $n^\circ \text{C} = n^\circ \text{F}$. It follows that $\frac{9}{5}n + 32 = \frac{5}{9}(n - 32)$. Clearing of fractions and solving for n , gives $n = -40$. The preceding relations are illustrated in Fig. 1, where the three scales are set side by side for comparison, their actual readings being shown at the usually accepted room temperature of 20°C .

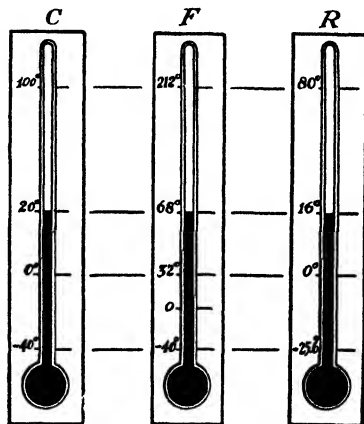


Fig. 1.

180. The maximum and minimum thermometer. There are various methods for recording or registering temperatures over a period of time, but the type seen most commonly is a thermometer designed to register the highest and lowest temperatures experienced since the instrument was last "set."

In the form invented by Six, the large bulb G shown in Fig. 2 is filled with a highly expanding fluid like phenol, or glycerine, and this takes the place of the bulb filled with mercury of the usual thermometer. Its expansion forces the mercury thread, which serves only as an index, down at a and up at b by equal amounts. Above b is another column of phenol or glycerine which partly fills the small bulb. The portion S of this bulb is partly exhausted, but retains enough air to act as a spring forcing the liquid downward when G

contracts. Thus the combined effect of the weight of the column g and the gas tension in S holds the mercury thread always in contact

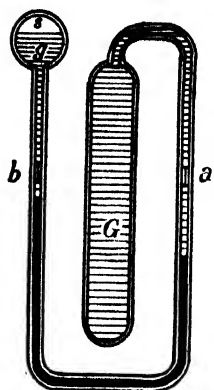


Fig. 2.

with the other liquid column in the right-hand branch of the tube, even though it contracts sufficiently to bring the right-hand end of the thread of mercury above the level of the left hand end. Above each end of the mercury column are small iron wires flanged at each end, so as not quite to fit the tube. They are readily pushed upward by the mercury, but stick at the highest point reached, and the glycerine or phenol then flows past them as the mercury recedes. The index at b is thus left at the maximum temperature reading attained, while the a index records the minimum. In order to prepare the thermometer for a new reading, the two iron markers are

drawn down to the mercury by a magnet applied from outside the glass tube.

SUPPLEMENTARY READING

T. Preston, *The Theory of Heat* (Chap. 1, sec. 1), Macmillan, 1894.

CHAPTER 14

Thermal Expansion

181. Linear expansion. As has been stated, the volume of most substances increases with rise of temperature, but in the case of an elongated solid like a wire, we may confine our attention to the change in length. Although the cross section increases also, this is of little practical interest, but the linear expansion of iron cables supporting suspension bridges, iron beams, girders, and steel rails is so great from winter to summer that it must be allowed for in the design of many structures.

As an example, the cable of a suspension bridge a quarter of a mile long expands about seven inches between freezing and 100° F. This would result in serious damage if the bridge were not sufficiently flexible. Rails of a railroad track must be slightly separated at their ends, for if they were not, the tendency to lengthen in warm weather would cause them to buckle and throw them out of alignment.

182. Coefficient of expansion. It is obvious that anything which tends to change the length of a bar or wire, such as tension or temperature, produces a greater effect on a long bar than on a short one. Such changes indeed are strictly proportional to the length. In the case of thermal expansion this change is also very nearly proportional to the change of temperature. Therefore $\Delta L \propto L\Delta t$ where ΔL is the change in length ($L_2 - L_1$) corresponding to the change in temperature Δt , or $t_2 - t_1$. This variation becomes an equation by introducing the constant of proportionality α , so that the equation $\Delta L = \alpha L\Delta t$ expresses the law of thermal expansion in a very compact form. The constant α is called the coefficient of linear expansion, and since $\alpha = \Delta L/L\Delta t = (L_2 - L_1)/L_1(t_2 - t_1)$, it may be defined as the *change in length per unit length per degree change in temperature*.

This last expression may be changed to a very convenient form by taking the initial temperature t_1 as zero, dropping the subscript of t_2 , setting $L_1 = L_0$, and $L_2 = L_t$. Then $\alpha = (L_t - L_0)/L_0 t$, and $L_t = L_0(1 + \alpha t)$. The graph of this equation is a straight inclined

line intersecting the L axis at L_0 , as in Fig. 3. But if plotted from very refined measurements, the experimental L - t curve may slope slightly upward, as indicated by the dotted line. This is because the

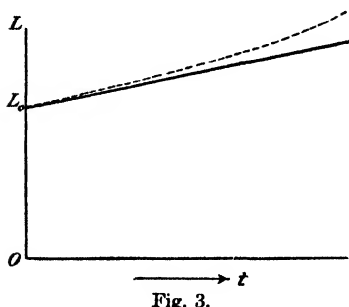


Fig. 3.

coefficient of linear expansion generally increases slightly as the temperature rises.

183. Practical applications. Although frequently a drawback, linear thermal expansion may be of great value in various processes and mechanisms. Shrinking the iron tire on a cart wheel has already been mentioned. Rivets are "headed" with blows of a

hammer when red hot, and as they shrink on cooling, they draw the plates of a boiler or girder tightly together.

When it is desired to magnify the effect of temperature, two strips of different metals having different coefficients of expansion are used. These are riveted together, as shown in Fig. 4, and if one end is clamped firmly, the other end moves in the direction of the arrow with rising temperature, provided the inner strip expands more than the outer one. Copper has a coefficient of expansion of 16.7×10^{-6} per degree centigrade, while the coefficient of wrought iron is only 11.9×10^{-6} . If such a differential device is made of these metals, with copper inside, the free end moves outward through a much greater distance per degree than the actual change in length of either bar, and may be used to drive a needle over a dial, as in the case of the dial thermometers in common use.

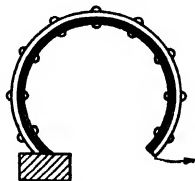


Fig. 4.

The balance wheel of a watch is a similar device, but here the metal of higher expansion is outside, and the inward bend of the free end offsets the increased diameter of the wheel caused by rising temperature. This increase would result in slowing the watch down because of its greater moment of inertia, were it not for the inward swing of the free ends of the semicircular rims, as shown in Fig. 5.



Fig. 5

In the "gridiron" pendulum of some clocks, the differential expansion is made use of to neutralize the lowering of the bob with rising temperature, and consequent lengthening of its period. An examination of Fig. 6 will show that on

either side are three rods (including the central one) whose expansion tends to lower the bob, and two rods (light lines) tending to raise it. If the metals used have expansion coefficients in the ratio of 2 : 3, then the pendulum maintains the same length at all temperatures.

184. Surface and volume expansion. A rectangular surface whose sides are a and b , may be thought of as made up of two systems of rods at right angles to each other. When heated, the rods parallel to the side a expand according to the relation $a_t = a_0(1 + \alpha t)$, and the expansion of those parallel to b is given by $b_t = b_0(1 + \alpha t)$. The area at t degrees is then $A_t = a_t b_t = a_0 b_0(1 + 2\alpha t + \alpha^2 t^2)$. But as α is very small, the term involving its square is negligible (unless t is excessively high) and may be dropped, so that the area equation becomes

$$A_t = A_0(1 + 2\alpha t),$$

where 2α is the coefficient of area expansion.

Similarly a rectangular parallelepiped whose edges are a , b , and c , has a volume at t given by

$$V_t = V_0(1 + 3\alpha t + 3\alpha^2 t^2 + \alpha^3 t^3),$$

in which the terms involving the higher powers of α may ordinarily be dropped, so that

$$V_t = V_0(1 + 3\alpha t) = V_0(1 + \beta t),$$

where β , the coefficient of volume expansion, is approximately three times that of linear expansion.

Although this has been proved only in the special case of a very regular solid, it is true for any other figure, because all volumes depend upon triple products of linear dimensions, so that the factor 3 must enter into the calculation of their expansion.

185. Expansion of mercury. In the case of any fluid, linear and surface expansion are meaningless, and we have only the volume coefficient β to consider. The importance of mercury as a thermometric fluid makes it necessary to know its behavior when heated, with unusual precision. This has been studied by various observers, but the classic measurement was made by Regnault, who improved upon a similar method by Dulong and Petit. The essential features of the apparatus are shown in Fig. 7 (a). Two iron tubes T contain mercury and pass up through the tanks A and B . The former is filled with oil, which is heated from below to any desired temperature

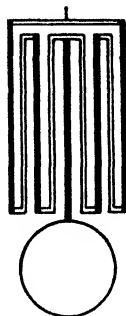


Fig. 6.

and constantly stirred to insure uniformity, while an air thermometer (not shown in the sketch) having a bulb as long as the tube, measures the temperature. The iron tube in *B* is surrounded by circulating cold water whose temperature at several levels is measured by ordinary thermometers. The mercury column is interrupted in glass

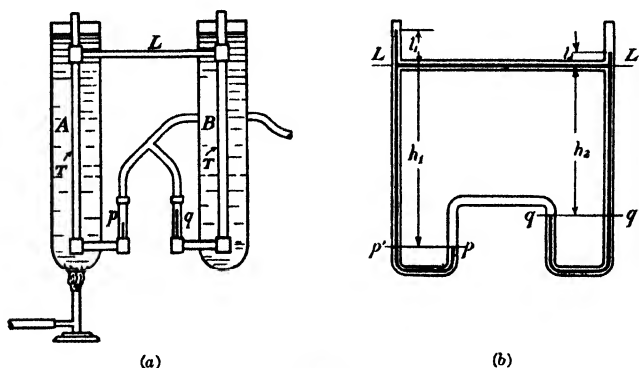


Fig. 7.

tubes at *p* and *q*, and compressed air, in a drum not shown, maintains the pressure necessary to support the differences of level $h + l$, indicated in Fig. 7 (b). The communicating tube *LL* serves to equalize the pressure at that level so that the columns l_1 and l_2 are in equilibrium. The upward pressures within the main columns at p' and q' are those of the compressed air acting downward at the corresponding levels *p* and *q*. These are balanced by the warm less-dense mercury column of height $h_1 + l_1$, and the cold denser column of height $h_2 + l_2$ respectively. The pressures due to these columns, $(h_1 + l_1)d_1g$ and $(h_2 + l_2)d_2g$ are therefore equal; also $l_1d_1g = l_2d_2g$, because of the equalizing tube *L*. Therefore $h_1d_1g = h_2d_2g$, or $h_1d_1 = h_2d_2$. But since density varies inversely as the volume of a given mass, the equation $V_t = V_0(1 + \beta t)$ may be written $d_t = d_0/(1 + \beta t)$; therefore $d_1 = d_0/(1 + \beta t_1)$ and $d_2 = d_0/(1 + \beta t_2)$. Then substituting for d_1 and d_2 in $h_1d_1 = h_2d_2$, we obtain

$$\frac{h_1d_0}{1 + \beta t_1} = \frac{h_2d_0}{1 + \beta t_2}$$

and

$$\beta = \frac{h_1 - h_2}{h_2t_1 - h_1t_2}.$$

This makes it possible to calculate the coefficient of volume expansion in terms of easily measured quantities, and within any chosen temperature interval.

Regnault's observations showed that β was not a constant, but could be expressed in terms of ascending powers of t . More recent observations, by Chappuis in 1907, show that the average value of β between 0°C and 100°C is given by

$$\beta = 1.8169 \times 10^{-4} - 2.951 \times 10^{-9} t + 1.15 \times 10^{-10} t^2.$$

Coefficient of Expansion (centigrade)

Solids	α (linear)	Liquids	β (volume)
Aluminum.....	25.5×10^{-6}	Water (5° to 10°)...	5.3×10^{-5}
Copper.....	16.7 "	Water (10° to 20°)..	15.0 "
Iron (cast).....	10.2 "	Water (20° to 40°)..	30.2 "
Iron (wrought).....	11.9 "	Alcohol.....	110 "
Steel.....	10.5 to 11.6×10^{-6}	Ether.....	163 "
Platinum.....	8.9×10^{-6}	Glycerine.....	50 "
Brass.....	18.9 "	Mercury.....	18.2 "
Glass (flint).....	7.8 "	Paraffin oil.....	90 "
Glass (soft).....	8.5 "	Sulphuric acid (concentrated)...	57 "

186. Expansion of water. Water, unlike most liquids, has a maximum density above its freezing point. Starting at zero, it contracts when heated, until the temperature reaches approximately 4°C (more exactly, 3.98°), when its density is taken as unity from the definition of the kilogram. Above 4° , water expands and becomes less dense, till at 100° its value is $0.9584 \text{ g per cm}^3$.

This anomalous behavior of water under 4° has some very important consequences. If a beaker of water is placed in a freezing mixture, and thermometers are inserted at different levels, it will be found that the lower layers soon reach 4° , while the surface temperature is still much warmer. This is because the water, as it cools around the sides and at the surface, becomes denser, settles to the bottom, and displaces the warmer water there. This rises to the surface, in turn becomes chilled, and then descends. Such a process occurs also in freshwater ponds and lakes, and if the cold weather continues long enough, the process of circulation described above ultimately results in lowering the temperature of the whole mass to 4° . As soon as this condition is reached, the upper layer no longer sinks on being cooled, but growing less dense, remains on top and ultimately freezes over. Thus ice, which is much less dense than water at 0° , not only forms on the surface, but floats there when formed.

This behavior of ponds accounts for the fact that in winter shallow ponds freeze first, and deeper ponds later and later in accordance with their depth. Some are so deep that they never freeze, because the cold weather does not last long enough to allow them to reach a uniform temperature of 4° , which is the necessary condition to be reached before the top layer can fall to zero.

Under certain conditions, to be explained later, water may be obtained both below zero and above 100° . Observations on its density in these regions show that it goes on expanding in both directions,

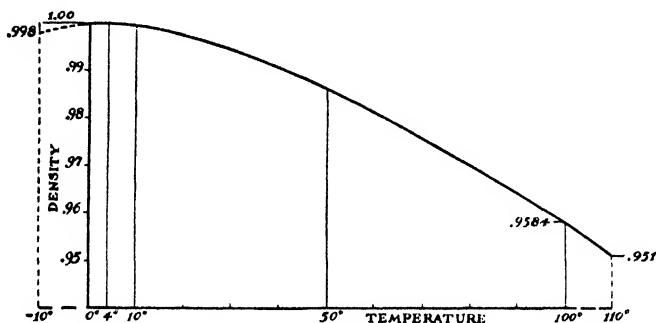


Fig. 8.

so that the curve of density plotted against temperature is convex upward with a maximum at 4° , and slopes off in both directions from this point, as shown in Fig. 8.

187. Expansion of gases. All gases when heated tend to expand, but as their volume at any temperature depends upon the pressure, it is necessary that the latter be kept constant if we are to obtain any rational observations of this effect, or if the coefficient of expansion is to be measured.

The first really accurate observations of the expansion of gases at constant pressure were made by Regnault, who used an apparatus shown in its essentials in Fig. 9. The gas to be examined is introduced from *F* into the bulb *C*, of large volume compared to *A*, by means of a three-way cock at *D*. The mercury, indicated by heavy shading, nearly fills the tube *A*, which is surrounded by water at a constant temperature. The bulb *C* is then cooled to some known temperature by immersing it in water or melting ice, and the cock *D* turned so as to shut off the supply from *F* and connect *C* with the tube *A*. The gas in *C* is then under atmospheric pressure *P* acting on the free surface of the mercury in *B*, plus an added pressure hdg

due to the difference of level h between the two columns. After taking this observation, the water surrounding C is warmed to some higher temperature and the gas in it expands, forcing down the level of the mercury in A , and forcing up the level in B . But by drawing off some mercury through the tap E , the original difference of level h can be restored, with the mercury in both A and B at lower levels than before. As the volume of C is known, as well as that of the connecting tube and the diameter of A , it is easy to compute both the original volume occupied by the gas, and the new volume when the mercury has fallen to the lower level LL in the tube A . Since the total pressure on the gas ($P + hdg$) is the same as before, the increase of volume for a given rise of temperature at constant pressure is readily determined.

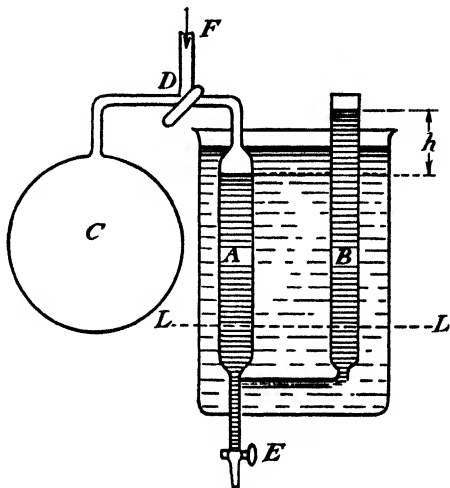


Fig. 9.

188. Coefficient of expansion. This quantity is found to vary somewhat at different pressures, and especially so in the case of such gases as carbon dioxide, whose deviation from Boyle's law is particularly marked. At atmospheric pressure the coefficient of volume expansion, usually designated by α , is 0.00367 per degree centigrade for air, and 0.00371 for CO_2 , but at about 250 cm of mercury, these values are 0.00369 and 0.00384 respectively. Hydrogen, being a more nearly ideal gas than either of these, varies between 0.0036613 and 0.0036616 under the same conditions, so that we may consider 0.00366, at all pressures, as a probable value for the coefficient of expansion of an ideal gas at constant pressure.

189. Charles's law. The French physicist Jacques Charles (1746–1823), through his interest in ballooning, seems to have been the first to notice the fact just stated—that all the common gases have about the same expansion coefficient. But Gay-Lussac, in 1802, made the first careful observations of this quantity, so that the law that *gases under constant pressure expand in equal proportion for equal increases*

in temperature is often called the law of Gay-Lussac as well as the law of Charles.

Still later Regnault made the observations already described, and found that the increase in volume per degree centigrade was $0.00366 = \frac{1}{273}$ of the original volume. We may therefore restate Charles's law as follows: *An ideal gas expands $\frac{1}{273}$ of its volume when heated through one degree centigrade from zero at constant pressure.* This may be expressed as an equation like that relating to the volume expansion of liquids and solids, but using α instead of β , as is customary with gases which have no linear expansion. Then

$$v_t = v_0(1 + \alpha t), \quad (1)$$

where the small letter v_t means the *specific* volume under some fixed pressure at the temperature t degrees centigrade, and v_0 means the *specific* volume under the same pressure at 0° , while α is the coefficient of expansion at constant pressure.

If the expansion does not start at 0° C, the relative change in volume deviates considerably from 0.00366 per degree when the initial temperature is very high. Thus if t_1 and t_2 represent the initial and final temperatures, equation (1) gives us

$$v_1 = v_0(1 + \alpha t_1) \quad (2)$$

$$\text{and} \quad v_2 = v_0(1 + \alpha t_2). \quad (3)$$

Dividing (3) by (2), we obtain

$$v_2 = v_1(1 + \alpha(t_2 - t_1) - \alpha^2 t_1 t_2 + \alpha^2 t_1^2 - \dots). \quad (4)$$

If t_1 is not far above zero, the terms in the ascending powers of the small quantity α are negligible, and (4) becomes

$$v_2 = v_1(1 + \alpha \Delta t) \quad (5)$$

where Δt is the temperature range. This is identical with (1). But if t_1 is large, the terms in α^2 , and so forth, must be considered.

SUPPLEMENTARY READING

E. Griffiths, *Methods of Measuring Temperature*, C. Griffin, London, 1925.

PROBLEMS

1. Convert 70° , 84° , 98° , 110° Fahrenheit to centigrade. *Ans.* $21^\circ 1$; $28^\circ 9$; $36^\circ 7$; $43^\circ 3$.

2. Convert 4° , 15° , 40° , 86° centigrade to Fahrenheit. *Ans.* $39^\circ 2$, 59° , 104° , $186^\circ 8$.

3. A copper bar 2.45 m long at 20°C is heated to 100°C . What is its increase in length? *Ans.* 0.327 cm.

4. A wrought iron tire whose inner diameter is exactly 70 cm at 0°C is to be shrunk on a wheel whose diameter is 0.666 cm too large. To what temperature must the tire be heated? *Ans.* 800°C .

5. The steel rails of a trolley line have a coefficient of expansion of 11×10^{-6} per degree centigrade and are 28 ft. long. How much space should be left between them when the temperature is 30°F , if their ends are just to touch at 120°F ? *Ans.* 0.18 in.

6. What is the change in volume of a brass kilogram weight which is heated from 20° to 100° , if its density at 20° is 8.4 g/cm^3 ? *Ans.* 0.54 cm^3 .

7. Calculate the volume of a flint glass flask which contains exactly 100 cm^3 at 20° , if it is heated to 80° . *Ans.* 100.14 cm^3 .

8. If the flask in Problem 7 had been filled with mercury, how many g of the liquid would have been driven out as a result of the expansion of both liquid and flask? *Ans.* 12.9 g.

* 9. A clock with a metal pendulum keeps accurate time at 0°C . Show that if the temperature is 35°C the number of seconds it loses in a day is given by $86,400 (1 - 1/\sqrt{1 + 35\alpha})$.

10. Calculate the specific volume of aluminum at 150° if its density at 20° is 2.65 g/cm^3 . *Ans.* $0.381 \text{ cm}^3/\text{g}$.

* 11. The brass scale beside a barometer is correct at 0°C . What is the actual height of the barometer when the reading is 77.40 cm at 30°C ? What would it have read at 0°C ? *Ans.* 77.44 cm; 77.02 cm.

CHAPTER 15

Ideal Gases

190. Change of pressures with constant volume. Regnault also studied this effect, using an apparatus similar to the one described in the last chapter, except that the mercury in the tube *A* (Fig. 9) is always brought back to its original level by pouring some mercury into the open end of *B*. The pressure, as before, is given by $P + hdg$, where h is now a variable distance increasing with the temperature, and d is the density of mercury. These observations show that the coefficient of change of pressure at constant volume is constant for ideal gases, and is the same as the constant-pressure coefficient, but these coefficients differ slightly in the case of real gases such as air or hydrogen. In the case of air, $\alpha_p = 0.003671$, while $\alpha_v = 0.003665$. The equation expressing the relation between pressure and temperature at constant volume, is similar to that of volume expansion. It reads

$$p_t = p_0(1 + \alpha_v t),$$

where p_0 is the pressure at 0° , and p_t is the pressure at the temperature t° centigrade. This is sometimes called Charles's law, although Charles's law is really the law of volume expansion:

$$v_t = v_0(1 + \alpha_p t).$$

191. Combined laws of Boyle and Charles. If a gram of gas is heated t degrees at the constant specific volume v_0 , the product of pressure and volume at this new temperature is $p_t v_0 = p_0 v_0(1 + \alpha_p t)$. Also, if heated to the same temperature at constant pressure p_0 , the product is $p_0 v_t = p_0 v_0(1 + \alpha_v t)$. In either case the gas is brought to the same temperature; therefore the product of pressure and volume must be the same at that temperature, according to Boyle's law. Then

$$p_t v_0 = p_0 v_t$$

$$\text{and} \quad p_0 v_0(1 + \alpha_v t) = p_0 v_0(1 + \alpha_p t),$$

whence

$$\alpha_v = \alpha_p.$$

This means that a gas which obeys Boyle's law exactly has the same coefficients for constant-pressure and constant-volume heating, as was demonstrated experimentally by Regnault.

Now let the condition of a gas be changed from $p_0 v_0$ at t° to any pressure p and the corresponding volume v at the same temperature. Then $p_0 v_0 = pv$, and

where α is either α_v or α_p , since they are equal, and the subscript is unnecessary. This equation combines the laws of Boyle and Charles, and enables us to calculate specific volumes or pressures at any assigned temperature.

192. The gas thermometer. The apparatus already described for determining the coefficients of expansion of a gas may be used to measure the temperature when α is known. If used as in the determination of α_p , the instrument is called a constant-pressure gas thermometer, and if used as in finding α_v , it is the constant-volume gas thermometer. The former is much less desirable than the latter, because when operating at constant pressure, the volume steadily increases, and it is very difficult to maintain the gas at a uniform temperature. This is because it expands more and more into the tube *A* (Fig. 9), where it is not at the same temperature as in the bulb *C*.

When the instrument is used with constant volume, the difficulty just referred to is almost wholly eliminated, provided the connecting tube is of very small bore, and the container *C* is large; because with the mercury level in *A* maintained near the top of the tube, only a very small amount of gas can be at a different temperature from the main body. This arrangement, when filled with such a gas as hydrogen, whose coefficient of expansion is practically constant at all usual pressures, is the most nearly ideal form of thermometer known. If an absolutely perfect gas were available, the changes in pressure would be an exact index of the temperature, since these changes can be determined with high precision in terms of the height of the barometer, and the difference of level of the mercury in *A* and *B*.

When measured in this way, temperature is often designated by θ , to differentiate it from t as measured by a mercurial thermometer. As this thermometer is laid off in 100 *equal* divisions between the freezing and boiling points of water, it has a scale based on the incorrect assumption that mercury expands uniformly over that range.

To determine θ from the pressure, we need only solve the equation $p = p_0(1 + \alpha\theta)$ for θ , giving

$$\theta = \frac{p - p_0}{p_0 \alpha}, \quad (1)$$

where p is the pressure when the bulb is heated to θ° , and p_0 is the pressure of the gas when the bulb is immersed in melting ice. But

this solution assumes an accurate knowledge of α for the gas employed, and since α is slightly different for different gases and is affected by impurities, it is better to eliminate it by a second observation, which consists in obtaining its value for the gas in question. This is done by immersing the bulb in boiling water whose temperature θ is 100° and observing the resulting pressure p_{100} .

We may then write

$$100 = \frac{p_{100} - p_0}{p_0 \alpha}. \quad (2)$$

Dividing equation (1) by (2), we eliminate α and obtain

$$\theta = 100 \left(\frac{p - p_0}{p_{100} - p_0} \right), \quad (3)$$

from which any temperature may be determined in terms of the corresponding pressure p and the pressures observed at the freezing and boiling points of water.

193. Absolute zero. If the gas used in a constant-volume gas thermometer is progressively cooled, the pressure falls by nearly equal amounts for each degree change of temperature. If we assume the gas to be ideal at all temperatures, this decrease of pressure would continue at an exactly uniform rate until there was no pressure left to measure. The temperature at which such a hypothetical gas would cease to exert any pressure at all is known as the **absolute zero**, and its position on the centigrade scale is easily found by setting $p = 0$ in the equation expressing the variation of p with t at constant volume. Then

$$p = 0 = p_0(1 + \alpha t).$$

But as p_0 , the pressure at the freezing point, is not zero, $(1 + \alpha t)$ must be zero. Then taking for α the average value 0.00366, we have

$$t = -\frac{1}{\alpha} = -\frac{1}{0.00366} = -273.2 \text{ approximately.}$$

This temperature has never been actually attained, although the Dutch physicist Kamerlingh Onnes of Leyden in 1921 reached the temperature of 0.82° above the absolute zero by evaporating liquid helium under greatly reduced pressure. Still more recently (1933), Professor Giauque of the University of California, operating with magnetic fields, reached -272.83°C , which is only a third of a degree above the absolute zero. Still lower temperatures, down to 0.0044° above absolute zero, have since been announced, but their exact value is somewhat uncertain.

194. The absolute or Kelvin scale of temperature. This scale starts at the absolute zero usually taken as -273°C (approximately -460°F), although it is really a little lower. Then the freezing point of water is at $+273^{\circ}$, and the boiling point at $+373^{\circ}$ Kelvin. Temperatures measured on this scale are indicated by T instead of t or θ , while K or A is placed after the number of degrees, as 150°K , or 150°A . It is understood that temperatures thus indicated are independent of the properties of any particular thermometric medium such as mercury, or even a nearly perfect gas like hydrogen. That the degree can be defined in this way was shown by Lord Kelvin, whose absolute scale is defined in terms of the energy supplied to a series of ideal engines, each operating on the heat rejected by its predecessor.

195. The gas law. The combined law of Boyle and Charles (equation (1) Article 191) may be expressed in terms of the absolute scale of temperature by simply substituting $T - 273^{\circ}$ for t . But because α equals the reciprocal of 273, very nearly, in the case of an ideal gas, $T - 273^{\circ} = T - 1/\alpha$.

Then

$$\begin{aligned}pv &= p_0 v_0 [1 + \alpha(T - 1/\alpha)], \\ &= p_0 v_0 \alpha T = p_0 v_0 T / 273, \\ &= rT, \text{ where } r = p_0 v_0 / 273.\end{aligned}$$

Since p_0 and v_0 are constants, r is also constant, but its numerical value depends upon what we mean by p_0 and v_0 . In the preceding paragraphs it was necessary to consider them only as pressure and specific volume at zero centigrade, but we may have any pressure at this temperature, and the volume at zero depends upon the pressure. It is therefore necessary in this definition of r to be more specific, and p_0 is understood to mean standard atmospheric pressure (hitherto denoted by P) which is the pressure just able to support a column of mercury 760 mm high at 45° latitude and at 0°C . The volume v_0 means specific volume under standard atmospheric pressure at the same temperature. As v_0 is different for different gases, r thus defined is a constant only as regards a particular gas. In the case of nitrogen, for instance, if measured in absolute units, $p_0 = 1,013,200$ dynes/cm², and if the specific volume is measured in cm³ per gram, $v_0 = 797.32$; therefore

$$r = \frac{1,013,200 \times 797.32}{273} = 2,959,138 \text{ ergs per gram per degree.}$$

196. Problems involving the gas law. If a problem concerning nitrogen is to be solved, the value of r just obtained may be used

provided we express p in dynes/cm² and v in cm³ per gram, while the temperature must be expressed in the absolute scale. Then any one of these three variables may be found from a knowledge of the other two by substituting in $pv = rT$. Or we might calculate r for any other choice of pressure or specific volume units, and then use it in the same formula in connection with the units chosen.

But fortunately it is rarely necessary to compute r . Most problems involving the three variables of the gas law concern the gas in two different conditions: $p_1v_1 = rT_1$, and $p_2v_2 = rT_2$, where the subscripts indicate the initial and final states of the gas. Then dividing the first equation by the second, we eliminate r , and obtain the very useful expression

$$\frac{p_1v_1}{p_2v_2} = \frac{T_1}{T_2}.$$

Here are six variables of which any five must be given in order that the sixth may be calculated. As an illustration of this type of problem, suppose 6 liters of air at atmospheric pressure and 20° C are compressed to 2 liters, with a rise of temperature to 30° C. Required, the final pressure p_2 . Since the pressure appears as a ratio, p_1/p_2 , it may be measured in any unit, as atmospheres, provided both of the pressures are measured in the same way. Therefore $p_1 = 1$, the ratio of specific volumes is 6:2, $T_1 = 273 + 20^\circ = 293^\circ$, and $T_2 = 303^\circ$. So that

$$p_2 = \frac{p_1v_1T_2}{v_2T_1} = \frac{1 \times 6 \times 303}{2 \times 293} = 3.1 + \text{atmospheres.}$$

197. The molecular "hypothesis." It has long been considered highly probable that matter is made up of discrete particles, each one being the smallest portion into which a given substance can be divided, and still retain its identity. This means a granular instead of a continuous structure, so that even such seemingly homogeneous materials as gold or water, when examined with sufficient minuteness, would appear as discontinuous aggregations of small bodies called molecules, all exactly alike.

The science of modern chemistry has been built upon such an assumption, in which the molecule of any chemical compound was regarded as a complex of the still more fundamental atoms of the elements. But even earlier than the discoveries of John Dalton (1766–1844) in this realm of chemical theory, Robert Boyle (1627–1691) and Robert Hooke (1635–1703) had suggested that heat could be explained by the agitation of the inner "parts" of a heated body,

which implied a belief in some sort of granular structure, and Hooke also accounted for the pressure of the atmosphere by supposing air to be made up of minute particles in rapid motion. Yet it was not until after the middle of the last century that Clausius, Clerk Maxwell, and Boltzmann worked out a complete molecular theory of gases. Their work on the "kinetic theory of gases" was purely mathematical, but experiments with cathode rays by Sir William Crookes, during the last decade of the century, began to give direct evidence of the corpuscular nature of matter, and in 1909, Jean Perrin of the University of Paris announced his remarkable proof of the molecular hypothesis based on the Brownian movements.

198. The Brownian movements. In 1827, a British botanist, Robert Brown, observed the curious irregular motions performed by small particles suspended in a liquid when viewed under the microscope. These particles are in a state of continual agitation and make a succession of abrupt motions, which, when plotted, form a confused network of no assignable pattern, as shown in Fig. 10. In order to obtain an emulsion containing sufficiently small particles to exhibit this effect satisfactorily, one may dissolve rosin in alcohol and pour a little of the solution into a glass of water. The milky precipitate which results consists of minute particles of rosin deprived by the water of the alcohol which had dissolved them, and their motions may be studied with the aid of a microscope, using light reflected from the particles when illuminated by a horizontal beam.

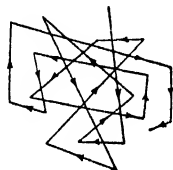


Fig. 10.

In 1879 Sir William Ramsay explained these mysterious movements of visible particles as caused by the bombardment of the vastly smaller molecules of the liquid which surrounded them. Each abrupt motion of the visible particle was due to the resultant of a great number of small impulses made upon it by the invisible molecules. These latter might then be inferred by their effect, very much as one might infer the existence of waves on the ocean when watching an ocean steamer pitching a long way off, although the waves themselves might be invisible at that distance.

Finally, during 1908 and 1909, Perrin conducted the experiments by which he was able, with the aid of theoretical ideas developed by Einstein, to calculate the number of atoms (or molecules) in a given mass of a substance. His results agreed so well with previous calculations based on totally different premises, that the molecular structure of matter has ever since been regarded as proved beyond any reasonable doubt.

199. Calculation of the pressure of a gas. The kinetic theory of gases developed by Clausius, Maxwell, and Boltzmann furnishes a method for calculating their pressure, based upon certain assumptions regarding their structure. These are: that gases are made up of discrete particles in rapid motion; that these behave like minute spheres of perfect elasticity; and that they are very small compared to the distances between them.

Such a swarm of particles moving at random inside an enclosure will have all sorts of velocities which increase with rising temperature, and they move through a great variety of distances between collisions with each other and with the walls of the enclosure. However, there must be an average velocity at a given temperature, and an average distance between collisions. These are known as the mean velocity, and the mean free path respectively, and their magnitude depends upon both the density and temperature of the gas.

Let us suppose the gas to be enclosed in a cubical box the length of whose edge is l centimeters. If the particles are assumed to be all of the same size and perfectly elastic, their mutual collisions will involve only interchanges of velocities whose net result on the total motion is zero, as was shown in the theory of collisions (Article 124). Thus we need consider only impacts with the walls of the box upon which this pressure is exerted. Then let us imagine a single molecule which moves back and forth between opposite walls in a direction normal to their planes. Let u represent its velocity supposed parallel to the X axis. Its momentum, mu , at each impact is reversed to $-mu$, since it rebounds with the same velocity in the opposite direction. Thus the *change* of momentum is $mu - (-mu) = 2mu$. But since the distance between opposite walls is l centimeters, the time t which elapses between impacts against the *same* face is numerically equal to $2l/u$. The time rate of change of momentum is obtained by dividing the change of momentum at each collision with a given face by the time between collisions, or $2mu \div 2l/u = mu^2/l$. But this rate of change of momentum measures the average force exerted by the vibrating molecule upon one face of the cube, and

$$F_{\text{av}} = \frac{mu^2}{l}. \quad (1)$$

Let n represent the number of molecules per cubic centimeter; then the number within the cube is nl^3 . Each of the six faces is bombarded by these nl^3 molecules moving in every conceivable direction. The velocities differ both in direction and magnitude, but may be resolved

into components parallel to X , Y , and Z axes, and denoted by u , v , and w . As the square of the velocities appears in equation (1), we must obtain the average square of the components, rather than square their averages; which in general gives a smaller result. These average squares, or *mean squares*, are represented by \bar{u}^2 , \bar{v}^2 , and \bar{w}^2 , and the mean square of their resultant, as was shown in equation (1), Article 20, is

$$\bar{c}^2 = \bar{u}^2 + \bar{v}^2 + \bar{w}^2. \quad (2)$$

The square root of this quantity, or $\sqrt{\bar{c}^2}$, is known as the "root mean square," r.m.s., and may be denoted by C . It is used in any calculation where the velocity is squared, as in kinetic energy. But in calculating mean momentum \bar{mu} , the ordinary mean velocity \bar{u} should be used.

As the motions of the gas molecules within the cube are absolutely random, the numerical values of the mean squares of their velocity components \bar{u}^2 , \bar{v}^2 , and \bar{w}^2 , are all equal; therefore in equation (2), $\bar{c}^2 = 3\bar{u}^2$, or

$$\bar{u}^2 = \frac{\bar{c}^2}{3} = \frac{C^2}{3}. \quad (3)$$

The average force exerted on a single face by all the molecules is nl^3 times that exerted by the single molecule of equation (1), and substituting $u^2 = C^2/3$ in (1), we obtain the average force exerted by all the molecules:

$$F_{av} = \frac{nmC^2l^2}{3}. \quad (4)$$

This force acts upon an area equal to that of the cube's face, or l^2 , so the pressure is obtained by dividing F_{av} by that area, giving $p = nmC^2/3$. But nm is the total mass of the gas within a cubic centimeter, which is the density d . Then

$$p = \frac{nmC^2}{3} = \frac{C^2d}{3},$$

or
$$pv = \frac{C^2}{3}, \quad (5)$$

where v is the specific volume, the reciprocal of density. But pv is constant at constant temperature; therefore C is constant at constant temperature, and is known if p and v (or d) are given. Its value for

hydrogen is found by substituting for d at 0°C , $8.99 \times 10^{-5} \text{ g/cm}^3$, and for p , $1.013 \times 10^6 \text{ dynes/cm}^2$ (normal atmosphere); then

$$C = \sqrt{\frac{3 \times 1.013 \times 10^6}{8.99 \times 10^{-5}}} = 1.839 \times 10^5 \text{ cm/sec.},$$

which is more than a mile per second.

The following table gives approximate values for C calculated as above at 0°C ; also values of the mean free path λ under standard conditions, and of the molecular diameter σ , and the molecular mass m .

Gas	C	λ	σ	m
Hydrogen..	$18.39 \times 10^4 \text{ cm/sec.}$	$18.3 \times 10^{-6} \text{ cm}$	$2.47 \times 10^{-8} \text{ cm}$	$3.1 \times 10^{-24} \text{ g}$
Nitrogen..	4.93×10^4	9.44×10^{-6}	3.50×10^{-8}	43.1×10^{-24}
Oxygen....	4.61×10^4	9.95×10^{-6}	3.39×10^{-8}	49.2×10^{-24}
Chlorine...	3.07×10^4	4.57×10^{-6}	4.96×10^{-8}	58.5×10^{-24}

200. Thermal energy as a function of absolute temperature. The equation $p = nmC^2/3$ may be written $p = \frac{2}{3}n \times \frac{1}{2}mC^2$. But $\frac{1}{2}mC^2$ is the average kinetic energy of a molecule within the enclosure, and if we denote this by W , the equation reads $p = \frac{2}{3}nW$. According to the gas law, $p = rT/v$; hence, equating these two expressions of the pressure and solving for W , we have the average energy of a molecule given by

$$W_{av} = \frac{3rT}{2vn} = \frac{3rdT}{2n}. \quad (1)$$

Since d is the mass per unit volume, or the mass of all the molecules within a unit cube, this quantity divided by the number of molecules gives the mass m of an individual molecule, or $d/n = m$, and

$$W_{av} = \frac{3}{2}rmT. \quad (2)$$

But $\frac{3}{2}rm$ is a constant for a given gas, so W varies as T , which means that the mean kinetic energy of the molecules of a gas is proportional to the absolute temperature.

201. Heat of compression. It was shown in Article 147 that it takes energy to compress a gas. The fact that this results in heating the gas may be shown from the kinetic theory just developed.

If the gas is enclosed in a cylinder fitted with a piston, and if this piston is pushed inward to compress the gas, the molecules striking its moving surface rebound with a higher velocity than if it had been still. This increased velocity results in an increased mean kinetic

energy, which denotes a higher temperature. Conversely, if the compressed gas is allowed to expand, pushing the piston outward, the rebounding molecules have their velocities reduced in proportion to its speed, and the temperature of the gas is lowered in consequence. Hence we may make the general statement that compression tends to heat a gas, while expansion against some opposition tends to cool it.

202. Avogadro's principle. As early as 1811, Amadeo Avogadro, an Italian physicist, advanced the hypothesis that *the total number of molecules per unit volume is the same for all gases at the same temperature and pressure*. But it was not till 1860 that this hypothesis was shown by Maxwell to be a consequence of the kinetic theory of gases, and therefore a "law" like that of Boyle. This may be proved as follows: Since $p = nmC^2/3$ we may express the equality of the pressures of two gases by

$$\frac{1}{3}n_1m_1C_1^2 = \frac{1}{3}n_2m_2C_2^2$$

But if their temperatures are equal, so are their mean kinetic energies, as we shall see in Article 204; therefore

$$m_1C_1^2 = m_2C_2^2, \text{ and } n_1 = n_2,$$

which is the principle usually known as Avogadro's law.

203. Avogadro's number. The **atomic weight**, a , of an element is the relative weight of an atom with respect to the weight of an atom of oxygen taken as exactly 16. Thus an atom of molybdenum, which weighs six times as much as one of oxygen, has an atomic weight of 96. An atom of carbon weighs three fourths as much; therefore its atomic weight is 12. This may be expressed by the proportion $a:16::m:m_o$, where a is the atomic weight of an element and m is the actual mass of one of its atoms, while m_o is the mass of an atom of oxygen.

Molecular weight, w , is the sum of the atomic weights of the atoms which make up the molecule. Carbon dioxide is a molecule having one carbon and two oxygen atoms, and its molecular weight is therefore $12 + (2 \times 16) = 44$.

Some molecules are made up of two atoms of the same element in chemical union, such as the gases hydrogen, oxygen, and chlorine, while mercury, whether in solid, liquid, or gaseous form is always monatomic. The molecular weight of oxygen gas is twice its atomic weight, or 32, while mercury, whose atomic weight is 200, has the same molecular weight. Therefore we may write a proportion similar to that for atomic weights, or $w_1/w_2 = m_1/m_2$, to express the fact that molecular weights, w , expressed in terms of oxygen as 32, are to each other as the actual masses, m , of the molecules themselves.

The mass of a compound equal numerically to the value of its molecular weight, as 32 grams of oxygen, is known as a **gram molecule**. If we divide this quantity by the mass, m , of each molecule, the quotient, N , is the number of molecules in a gram molecule. Now if the proportion between molecular weights and weights of molecules is transformed to read $w_1/m_1 = w_2/m_2$, each member of the proportion represents the quotient N ; for w is either molecular weight or the mass of a gram molecule by definition. A similar relation may be obtained between atomic weight (or a gram atom) and the mass of an atom, and the ratios $a_1/m_1 = a_2/m_2$ are also equal to N , since w is as much larger than a as the molecule is heavier than the atom. Therefore

$$\frac{a}{m_a} = \frac{w}{m_m} = N.$$

We may then make the general statement that *the number of atoms in a gram atom, or molecules in a gram molecule, is the same for all elements or chemical compounds*. This is known as **Avogadro's number** N , and has been found by a variety of methods to be 6.06×10^{23} . The most direct of these methods is that of Perrin using the Brownian movements, but because of the experimental difficulties involved, it is not the most accurate. Perrin obtained 6.85×10^{23} , which is considered a remarkably close agreement with the more reliable values obtained indirectly from more precise data.

204. Avogadro's principle and ideal gases. The gram molecules ("moles") of gases all have the same number of molecules. We also know that the volumes vw occupied by a *mole* are the same for all at the same pressure and temperature. If we multiply the gas law $pv = rT$ by w , it becomes $pvw = rwT$. But as vw is a constant for all gases when p and T are constant, rw must be a constant also. Therefore, denoting this general constant rw by R , we have

$$pv = \frac{RT}{w}. \quad (1)$$

When p is measured in dynes per cm^2 , and v is in cm^3 per gram, R equals 8.313×10^7 ergs or 8.313 joules per gram molecule per degree for all gases insofar as they may be regarded as "perfect." We may then calculate any of the three quantities p , v , or T for a gas if the other two are given, and its molecular weight is known.

As an illustration, the volume occupied by ten grams of oxygen under a pressure of two atmospheres at 20°C is calculated as follows: The specific volume is the required volume V divided by 10. The

pressure is $2 \times 1,013,200$ dynes/cm². The molecular weight is 32, and $T = 293^\circ$ K. Then $2,026,400 \times V/10 = 8.313 \times 10^7 \times 293/32$. Whence $V = 3.757$ liters, which agrees very closely with the value 3.755 based on actual observation.

In Article 200 we found that the mean kinetic energy of a molecule of an ideal gas was given by

$$W_{av} = \frac{3}{2}rmT. \quad (2)$$

The energy, W_m , of a gram molecule of such a gas is N times this value, or $\frac{3}{2}rmNT$. But $r = R/w$ by definition, and $m = w/N$; therefore equation (2) becomes

$$W_m = \frac{3}{2}RT, \quad (3)$$

which is a general expression for the thermal energy of a gram molecule of an ideal gas in terms of its temperature. Dividing (3) by N , we obtain $W_{av} = \frac{3}{2}RT/N$. This may be written

$$W_{av} = \frac{3}{2}kT, \quad (4)$$

where k (equal to R/N) is known as Boltzmann's constant. It equals 1.37×10^{-16} erg per degree, and may be thought of as the gas constant of a single molecule of an ideal gas. It is much used in the quantum theory to be discussed in a later chapter.

Equations (2), (3), and (4) are based on the assumption that the gas molecules are particles having three degrees of freedom. That is, they do not rotate but can be translated in three spatial dimensions. This is the case with monatomic gases like argon. But with diatomic gases like hydrogen and oxygen, two additional degrees of freedom are added because, like a dumbbell, they can rotate around each of two axes perpendicular to each other and to the line joining the atoms. Then the molecular energy becomes $W_{av} = \frac{5}{2}kT$. Molecules having three atoms have one more degree of freedom, making $W_{av} = \frac{6}{2}kT$, and in general each degree of freedom adds $kT/2$ to the molecular energy.

The pressure of an ideal gas may also be calculated in a generalized form as follows:

$$pv = RT/w.$$

But $d = nm$; therefore $v = 1/nm$, and

$$p = nmRT/w.$$

Substituting $w/m = N$, we obtain

$$p = \frac{nRT}{N},$$

or

$$p = nkT. \quad (5)$$

SUPPLEMENTARY READING

T. Preston, *The Theory of Heat* (Chap. 2, sec. 2), Macmillan, 1894.

Jean Perrin, *Atoms*, Van Nostrand, 1916.

Saha and Srivastava, *A Textbook of Heat* (Chap. 3), Indian Press, Allahabad, 1931.

PROBLEMS

1. Calculate the pressure in 12 g of hydrogen which occupy 1.8 l at a temperature of 80°C . (Consult table of densities, end of Chapter 9.) *Ans.* 95.9 atmospheres.

2. Calculate the mass of oxygen in a space of 25 l when the temperature is 30°C , and the pressure 0.9 atmosphere. *Ans.* 29 g.

3. The volume of a balloon is 200 cubic yards at atmospheric pressure and 70°F . What is its volume at an altitude where the pressure is 0.8 atmospheres and the temperature is 30°F , if the amount of gas remains the same? *Ans.* 231.1 cubic yards.

4. A cylinder contains 200 cm^3 of an ideal gas under a pressure of 56 cm of mercury at a temperature of 20°C . It is compressed to 40 cm^3 , when the pressure is found to be 300 cm. What is the final temperature? *Ans.* 41°C .

5. A mass of 250 g of air at 0°C and atmospheric pressure is compressed into a container of 50 l with a final temperature of 40°C . What is the resulting pressure? *Ans.* 4.4 atmospheres.

* 6. A piston weighing 8 kg and sliding without friction in a vertical cylinder 18 cm in diameter rests upon an air cushion which it has compressed to occupy a space 30 cm long. The air is then heated 200°C above its original temperature of 0° . Calculate the pressure below the piston (assuming normal atmosphere above), the volume after heating the air, and the work done. *Ans.* 1,044,100 dynes/cm²; 13,230 cm³; 584 joules.

7. Calculate the mass of air in a room $3 \times 4 \times 6\text{ m}$ at 30°C and a pressure of 72 cm of mercury. *Ans.* 79.47 kg.

8. The barometer reads 32 cm of mercury on the top of a high mountain, and the temperature is -20°C . What is the percentage of the density of the air to its density at sea level under normal conditions (0°C and 76 cm)? *Ans.* 45.4 per cent.

* 9. Calculate the value of the constant r for nitrogen and oxygen when pressures are measured in bars and specific volumes in cm³ per g. *Ans.* $r_o = 2.597 \times 10^6$; $r_n = 2.959 \times 10^6$.

* 10. Calculate the energy in a cubic centimeter of oxygen at 0°C and standard pressure. *Ans.* 0.253 joule.

CHAPTER 16

Heat Measurements

205. Quantity of heat. As was proved in the last chapter, temperature depends upon the mean kinetic energy of the molecules of a gas, but in solids and liquids the motion of the molecules must also determine their temperature in some way not yet explained. However, if we add up all the kinetic energies of all the molecules in a body, we arrive at a quantity very different from temperature, for the latter is not kinetic energy at all, but only depends upon or is determined by it. This sum of all the molecular kinetic energies constitutes *quantity* of heat, while temperature is only an index of the level at which this energy exists, just as a reservoir may contain a million gallons of water at a level either of one hundred or of two hundred feet above the town it supplies.

206. Units of heat quantity. Since heat has been fully demonstrated to be a form of energy, it might seem reasonable to measure it in ergs. Indeed, this can be done, as we shall see, but for practical reasons, a unit based directly upon purely thermal or *calorimetric* methods of measurement, is much more convenient.

In the c.g.s. system, this unit is the **calorie**, which is defined as *the amount of heat required to raise one gram of water from 15° to 16° centigrade*. This interval is chosen because it is equal to the average value per degree taken over the range from freezing to boiling. To distinguish this unit from a larger one defined as the heat required to raise the temperature of a *kilogram* of water one degree, the smaller unit is often called a small or gram calorie, as opposed to the large or kilogram calorie.

In the English system a similar unit exists known as the **British thermal unit**, or B.t.u. It is rarely used in working physical problems, but is common in power-plant engineering, and is defined simply as the amount of heat required to raise one pound of water one degree Fahrenheit. Therefore, since a pound equals 453.6 grams, and a degree Fahrenheit is 5/9 of a degree centigrade, one B.t.u. = $5/9 \times 453.6 = 252$ gram calories.

207. Thermal capacity. Everyone knows that it takes longer to heat some objects over a fire than others, and that those which take the longest to heat take longest to cool off again. It takes longer to heat a kettle full of water than one only half full, while the same amount of oil can be heated much more rapidly. This means that the quantity of heat required to produce a given change of temperature depends upon both the mass and the nature of the substance, or the heat required per degree rise of temperature varies as the mass of the substance times a constant whose value is determined by the nature of the substance heated. This may be expressed as

$$H = sm(t_2 - t_1), \quad (1)$$

where H is the heat in calories required to raise m grams of the substance from $t_1^\circ \text{C}$ to $t_2^\circ \text{C}$, and s is a characteristic constant known as the **specific heat** of the substance.

The product sm of the specific heat and the mass of a body consisting of a single substance is known as the **thermal capacity** of the body, because it is the heat in calories required to raise the body's temperature one degree, as is seen by setting $t_2 - t_1 = 1$ in equation (1). This quantity sm is also known as the **water equivalent** of the body for the following reason: If equal quantities of heat raise a body and a certain mass of water one degree in temperature, from (1)

$$H = s_1 m_1 \times 1^\circ = s_w m_w \times 1^\circ.$$

But the specific heat of water s_w is approximately unity at all ordinary temperatures; therefore

$$s_1 m_1 = m_w,$$

where m_w is the equivalent mass of water.

208. Measurement of specific heat. Specific heat is most easily measured by comparing the thermal capacity of any substance with that of water. This is accomplished by the "method of mixtures," when a certain mass of the substance at a known temperature is "mixed" with water at another known temperature, usually lower than the first. When two bodies at different temperatures act upon each other in this way, one loses heat with a fall of temperature, while the other gains it as its temperature rises, the final temperature of the mixture being somewhere between the two initial values. Moreover, as heat is a form of energy, it is indestructible, as will be explained later; therefore the heat lost by one substance (or body) is equal to that gained by the other.

If t_3° represents the final temperature of the mixture, then the body at the higher temperature falls from t_1° to t_3° , while the other rises from t_2° to t_3° . Then, equating heat lost by the one to heat gained by the other, we have $H_1 = H_2$, or

$$s_1 m_1 (t_1 - t_3) = s_2 m_2 (t_3 - t_2).$$

If the colder substance is water, then $s_2 = 1$, and the specific heat of the warmer body is given by

$$s_1 = \frac{m_2 (t_3 - t_2)}{m_1 (t_1 - t_3)},$$

from which s_1 may be computed in terms of easily measured quantities.

This equation enables us not only to determine the specific heat, but if s_1 is known, to find any one of the five other quantities if four of them are given. Thus the final temperature of a specified mixture may be calculated, or one of the two masses, or one of the two initial temperatures.

209. Mixtures of more than two substances: Since there is always a containing vessel, called a calorimeter (when a solid is immersed in a liquid), and other bodies necessary for accurate observation, such as a stirrer, the heat absorbed or given out by these bodies must be taken into account, as well as the fact that the liquid is not necessarily water.

In order to do this, the thermal capacity of each article must be added to one side of the equation or the other to allow for the amount of heat it absorbs or gives out. Thus if the container is of copper having a mass m_c , it absorbs $s_c m_c (t_3 - t_2)$ calories in being heated from t_2 to t_3 , where $s_c m_c$ is the water equivalent of the container.

We may then equate heat gained by the liquid, container, stirrer, and so forth, to heat lost by the bodies put into the liquid, obtaining $(s_1 m_1 + s_c m_c + s_s m_s + \dots)(t_3 - t_2) = (s_x m_x + s_y m_y + \dots)(t_1 - t_3)$. Then collecting similar terms under the Σ sign, we have $\Sigma(sm)_2 \times (t_3 - t_2) = \Sigma(sm)_1 \times (t_1 - t_3)$, where $\Sigma(sm)_2$ is the sum of the water equivalents of all the bodies heated from t_2° to t_3° and $\Sigma(sm)_1$ is the sum of the water equivalents of everything which was cooled from t_1° to t_3° . Any one of the various quantities represented, such as s_1 , s_x or t_1 , may then be found, provided all the rest are known.

210. Specific heats of solids and liquids. In general, liquids have higher specific heats than solids, water having the highest value of all.

It will be seen from the following table that some *metals* have values as low as three per cent of the specific heat of water, while only a few *liquids* (not metallic) have values as low as forty per cent of that

Liquids	t (°C)	s	Solids	Average t (°C)	s
Ethyl Alcohol.	0	0.547	Aluminum.	0-100	0.2114
Methyl Alcohol.	12	0.601	Copper.	0-100	0.091
Benzene.	10	0.34	Lead.	0-100	0.031
Ethyl Ether.	18	0.56	Platinum.	0-100	0.032
Glycerine.	18	0.58	Silver.	0-100	0.055
Toluene.	18	0.40	Brass.	0	0.090
Paraffin Oil.	20	0.51	Crown Glass.	10-50	0.16
Mercury.	20	0.0333	Flint Glass.	10-50	0.12
			Ice.	-21-(-1)	0.502

of water. The necessity for giving the temperatures in the above table arises from the fact that specific heats are not constant, but in almost all substances increase with the temperature according to the empirical formula $s = a + bt + ct^2 + \dots$

In the case of the various forms of carbon, the change of s with the temperature is quite rapid, so that at 200° C, Weber found that the specific heat of diamond was three times its value at 0°. But in the case of most solids, s increases quite slowly until near the melting point, when a rapid change occurs, giving a decidedly higher value as the liquid state is reached. Thus water in its solid state (ice) has a specific heat of about 0.5, while it is unity in the liquid state. The specific heat of lead changes from 0.034 to 0.040, while that of tin changes from 0.056 to 0.064 in becoming liquid.

In general, the specific heat of liquids increases with the temperature, though it may decrease at first, pass through a minimum value, and then increase as the boiling point is reached. The specific heat of mercury is 0.0335 at 0° C, 0.0327 at 100°, and continues to decrease slightly to a minimum value at 140°, after which it increases steadily.

211. Specific heat of water. Because of its great importance in calorimetric measurements, the specific heat of water at various temperatures has been determined with great precision. The results are shown approximately in the accompanying curve (Fig. 11), which is based on the calorie defined as the average thermal capacity of water over the range between freezing and boiling. There is a minimum of 0.9971 near 40°, and the specific heat is exactly 1 at 15.5 and again

near 70° , so that the total heat represented by the diagonally ruled area ($s > 1$) should equal the heat represented by the vertical shading ($s < 1$).

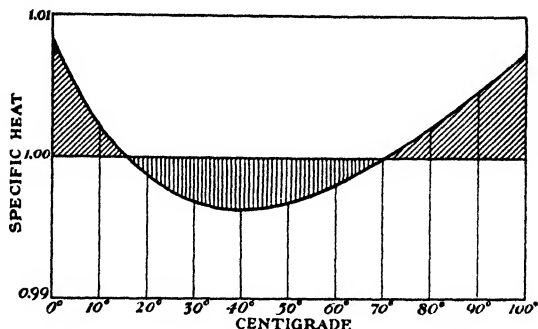


Fig. 11.

212. Law of Dulong and Petit. It is natural to inquire whether the specific heats of the elements follow any general law connected with their properties as atoms. Following an investigation with this in view, the French physicists, Dulong and Petit, in 1819 announced the law that *the product of the specific heat by the atomic weight is the same for all elementary solid substances*. This product averages 6.38 at room temperature for 32 substances, ranging between 5.7 and 6.76. But carbon, boron, and silicon (related elements) are exceptions. Carbon, in its allotropic form known as graphite, has a product of only 2.39. These exceptions, however, approach the normal atomic heat of 6 as their temperature rises, because their specific heats rise with the temperature. Diamond (a form of carbon) reaches this value at 980°C .

The meaning of this law is that the atoms of the elementary substances have nearly equal specific heats at ordinary temperatures. This can be shown as follows: If a is the atomic weight of an element, it is also the weight of a gram atom; therefore sa is the heat required to raise one gram atom through one degree. It is called the **atomic heat** of the element. This quantity is nearly constant, as we have seen, and if divided by Avogadro's number N , the quotient is the specific heat of the individual atom, which is therefore nearly constant also.

213. Specific heats of gases. When a body is heated under atmospheric, or any other pressure, its expansion involves work. This, however, is so small for solids and most liquids that its effect on the

specific heat may be ignored. But in the case of gases, the effect is very marked and results in a variable specific heat according to the method of heating. There are two especially important ways of heating a gas: one when the volume is kept constant by enclosing the gas in a rigid container, and the other when it is kept at constant pressure, as in Regnault's constant-pressure gas thermometer.

When a gas is heated at constant volume, the heat supplied to it does nothing but increase the mean kinetic energy of the molecules, and thereby raises its temperature. But when it is allowed to expand as a result of this increased thermal energy, it does work against the opposing constant pressure. Therefore the heat supplied is turned partly into mechanical work, and in order to produce the same change of temperature as before, a greater amount of heat must be supplied, and it follows that the specific heat s_p at constant pressure is greater than s_v at constant volume.

The ratio of the specific heats, usually represented by the Greek letter γ (gamma), is about 1.66 for gases like argon and mercury vapor whose molecules consist of a single atom. It is 1.4 for gases whose molecules are composed of two atoms, like oxygen and nitrogen, the chief constituents of air, while it falls to around 1.33 for gases whose molecules are made up of three atoms, and still lower for polyatomic vapors like that of alcohol. The value for air, determined with great care because of its obvious importance, is 1.4029.

214. Kinetic theory of the specific heats of gases. It was shown in Article 147 that when a gas expands through a volume Δv at constant pressure, the work done is $p\Delta v$. This may be written $W = p(v_2 - v_1)$, where v_1 and v_2 are the specific volumes before and after expansion, and W is the work done by a gram of the gas. Such an expansion implies a rise of temperature to maintain p constant; therefore, using equation (1) of Article 204, we have

$$W = p(v_2 - v_1) = R(T_2 - T_1)/w.$$

Then if the rise of temperature is one degree, $T_2 - T_1 = 1$ and

$$W_1 = p(v_2 - v_1) = R/w. \quad (1)$$

In Article 213 it was explained that the specific heat at constant pressure is larger than that at constant volume because of the external work done by the expanding gas; therefore if *one gram* of gas is heated *one degree* and expands against a constant pressure, the work done is due to the excess thermal energy represented by the difference of the specific heats measured in calories per gram per degree. This may be

converted into mechanical energy units by a factor J , to be more fully explained in Chapter 20, and we may write W_1 (ergs) = $J(s_p - s_v)$, where $J = 4.185 \times 10^7$ ergs per calorie. Combining this with (1) above, we have

$$J(s_p - s_v) = R/w. \quad (2)$$

If a gram molecule of a gas is heated within a rigid enclosure (v constant) from 0° K to T° K, no *external* work is done, but $ws_v T$ calories or $Jws_v T$ ergs of heat must have been supplied in the form of *internal* energy, shown by the rise of temperature. That is,

$$W_m = Js_v w T,$$

where W_m is the internal energy at T . But by equation (3) of Article 204, $W_m = \frac{3}{2}RT$. Therefore, equating these values of W_m , we obtain

$$Js_v = \frac{3}{2}R/w, \quad (3a)$$

$$\text{and} \quad s_v = 3r/2J, \text{ or } S_v = 3R/2J, \quad (3b)$$

where S denotes the specific heat of a gram molecule. Eliminating s_v between (2) and (3a), we obtain

$$Js_p = \frac{5}{2}R/w, \quad (4)$$

and dividing (4) by (3a), we have

$$s_p/s_v = \gamma = \frac{5}{3} = 1.66, \quad (5)$$

which was stated as an experimental fact in Article 213. Thus the kinetic theory of gases composed of ideal, perfectly elastic particles has led us to obtain the observed value of the ratio of the specific heats of a monatomic gas. This ratio for diatomic gases, calculated in a similar manner, is 1.4.

The specific heat S_v of a gram molecule of an ideal monatomic gas is readily computed by substituting in (3b) the numerical values of R (see Article 204), and of J (4.185 joules per calorie). The result is $(3 \times 8.313)/(2 \times 4.185) = 2.979$, which is a theoretical constant for all such gases. The calculated value of S_v for diatomic and polyatomic gas shows a steady increase with an increasing number of degrees of freedom.

In the case of those solids which may be regarded as monatomic, the atom is thought of as vibrating about a mean position of rest with a continual exchange between potential and kinetic energy as with a swinging pendulum. The mean values of these forms of energy are equal to each other in such systems. Therefore the total energy is twice the kinetic, or $W_m = 2(\frac{3}{2}RT) = 3RT$, and the

heat capacity of a gram atom, or the atomic heat, is twice that of a monatomic gas. It is given by $S_v = 3R/J = (3 \times 8.313)/4.185 = 5.96$, which is near the average of the values on which the law of Dulong and Petit was based.

Density and Specific Heats of Gases
(normal atmospheric pressure)

	Density (0°C) (grams per liter)	s_p (at 20°C)	s_p (at 20°C)
Air.....	1.293	0.2417	0.1724
Carbon Dioxide.....	1.977	0.2020	0.155
Hydrogen.....	0.0899	3.39	2.40
Nitrogen.....	1.2507	0.246	0.177
Oxygen.....	1.4290	0.220	0.157
Ammonia.....	0.7708	0.518	0.385

215. Quantum theory of specific heat. The classical theory of specific heat outlined above does not account for the experimental fact that specific heats decrease progressively as the temperature approaches the absolute zero. In order to explain the observed facts, Einstein made use of Planck's quantum theory. According to this theory (more fully discussed in Article 537) atoms or molecules can absorb or give up energy only in definite amounts called **quanta**. Thus the flow of energy is not continuous, but in steps, like a flight of stairs. On this assumption Einstein developed expressions for the average molecular energy and the specific heat of a gas at constant volume at a given temperature. Both expressions depend upon the quantity ϵ/kT where ϵ is the energy of a quantum and k is Boltzmann's constant R/N . If T is large, ϵ/kT is small for usual values of ϵ , and the average molecular energy approaches $kT/2$ for each degree of freedom in accordance with the kinetic theory, as explained in Article 204. Also when T is large, the specific heat of a monatomic gas at constant volume approaches $3r/2J$, as in equation (3b), Article 214. But if T is small and ϵ is constant, the calculated specific heat decreases with falling temperature, and becomes zero when $T = 0$, as indicated by experiment.

Einstein's theory further shows that when $T = \epsilon/k$, a gas having more than one atom per molecule begins to behave like a monatomic gas with a corresponding decrease of specific heat. This was shown to be the case by Eucken in 1912. He found that at -233°C the specific heat S_v of a gram molecule (2 grams) of hydrogen falls to 2.98, though its value at 20°C is $2 \times 2.40 = 4.80$. The temperature at which

this change takes place is called the **characteristic temperature**, denoted by Θ . Solids as well as gases have such a temperature when the molecular heat begins to fall below Dulong and Petit's average value of 6.38. The higher the characteristic temperature, the smaller the specific heat at lower temperatures. Diamond has an exceptionally high Θ , 1860° K, and at 20° C its specific heat S is only 1.4, while at -230° it is practically zero. With lead, on the other hand, Θ is 90° K, and its molecular heat at room temperature is $207 \times 0.031 = 6.42$, which agrees closely with Dulong and Petit's value.

Thus the quantum theory accounts for the observed facts much better than the older kinetic theory, though it is not very satisfactory near the absolute zero. A still more recent theory due to Debye works better at extremely low temperatures. The theory is partly based on the assumption that when T is very small the specific heat varies as T^3 . This is known as "Debye's T^3 Law."

SUPPLEMENTARY READING

C. H. Draper, *Heat* (Chap. 7), Blackie & Sons, London, 1911.

PROBLEMS

1. How many calories are required to heat 45 g of mercury from -20° C to $+60^\circ$ C? *Ans.* 119.9 calories.
2. A block of copper rises in temperature from 20° to 80° when it absorbs 4000 calories. How much does it weigh? *Ans.* 732.6 g.
3. A mass of 480 g of a solid substance at 100° C is placed in a bath of 200 cm³ of water at 20° C. The copper calorimeter which contains the water weighs 80 g. The final temperature of the mixture is 37.4° C. Calculate the specific heat of the substance. *Ans.* 0.12 calories per gram.
4. A mass of 300 g of lead at 100° C is plunged into a bath of 100 cm³ of glycerine contained in a 40 g aluminum calorimeter at 0° C. What is the resulting temperature? *Ans.* 10.2° C.
5. A piece of red-hot platinum weighing 25 g is plunged into 80 cm³ of water at 20° C in a crown-glass beaker weighing 90 g. The final temperature is 30° . What was the temperature of the platinum? *Ans.* 1210° C.
6. How much water at 96° C should be added to 200 cm³ of ethyl alcohol at 18° C to bring the temperature up to 60° C, allowing 14 g for the water equivalent of the vessel which contains the alcohol? *Ans.* 117 g.
7. A piece of brass, weighing 300 g, at 100° C, is dropped into 285 g of water at 10° C contained in a vessel weighing 150 g and of 0.1 specific heat. The final temperature is 17.3° C. Find the specific heat of this sample of brass. *Ans.* 0.088 calories per gram.

CHAPTER 17

Change of State

216. What change of state means. There are three so-called *states* of ordinary matter: solid, liquid, and gaseous. We have so far been concerned only with matter regarded as permanently in one of these three conditions. But as all the elements and most of the chemical compounds may exist in any one of the three states under suitable conditions, their behavior in passing from one state to another is of great importance. This process in general is called **change of state**, and may mean the change from solid to liquid (or vice versa), liquid to gas (or vice versa), and solid to gas or gas to solid without the appearance of the liquid state.

217. The melting point. When a substance changes from solid to liquid, the process is called **fusion**, or melting, while the reverse process is known as solidification or freezing. Solids of a crystalline character change to the liquid state at a very sharply defined temperature called the melting point, and some of these temperatures are known so accurately that they serve as "fixed points" by means of which thermometers may be calibrated over ranges quite outside the interval between the fixed points of freezing and boiling water.

Noncrystalline or amorphous (formless) solids do not melt at any definite temperature, but gradually soften and liquefy as they grow hotter, like glass, paraffin, gelatine, and so forth, so that it is impossible to say at just what temperature they became liquid.

Freezing of crystalline substances takes place at the same temperature as melting, provided they crystallize in the process. If they do not, as is the case with some substances normally crystalline in the solid state, their solidification is a gradual process at no definite temperature, as in the case of amorphous bodies.

Some of the more important melting points which may be used as standard temperatures are given in the following table:

Substance	Melting Point	Substance	Melting Point
Mercury.....	-38°87 C	Zinc.....	419°4 C
Water.....	0°	Antimony.....	630°
Tin.....	231°84	Sodium Chloride...	801°
Cadmium.....	320°9	Platinum.....	1770°

218. Change of volume during fusion. Most substances expand when they change from solid to liquid, or, conversely, contract in freezing, but the fact that ice floats with about 10 per cent of its bulk out of water shows that it must have expanded in the process of freezing and that water is an important exception to the general behavior. This expansion, which occurs in the case of only a few other substances, results in water pipes bursting when they freeze, and in the breaking up of rocks in winter by the formation of ice in minute cracks where water has penetrated them. Cast iron, bismuth, and the alloy known as type metal have the same property of expanding during solidification. This is most desirable in making castings, for it causes the molten metal to fill the mold completely during solidification. It shrinks afterward in the process of cooling, but still retains the exact form of the mold, though with smaller volume. To allow for this final decrease, the wooden models used in forming the molds for iron castings are built with the use of a "shrink rule" whose foot and inch divisions are larger than they should be, so that the final result is a casting reproducing the model in form and of the desired size.

The succession of volume changes just referred to may be shown graphically for the two types of substances, as in Fig. 12. Here (a) shows the behavior of the more

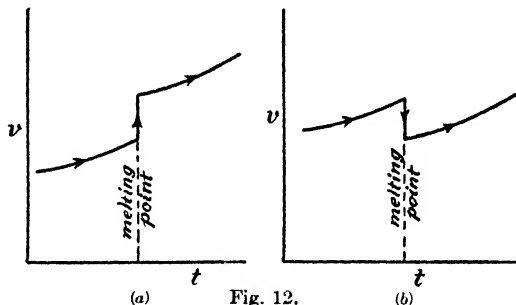


Fig. 12.

usual substances before, after, and during fusion, while (b) shows that of those which contract during fusion, such as water. In both cases, solid and liquid expand with rising temperature (except water between 0° and 4°), but the abrupt change during fusion is upward in one case, and downward in the other.

It is interesting to note that not only does the volume change in the process of melting, but the vapor pressure and electrical conductivity as well. The conductivity increases abruptly in those substances which contract in melting, and decreases in those which expand. The metal zinc, for instance, is more dense in the solid than in the liquid state, and has just half as much conductivity after it is melted.

219. Change of the melting point due to pressure. The exact temperature at which melting or freezing occurs depends upon the

pressure exerted upon the body during the process. The values given in the preceding table assume normal atmospheric pressure, but if the body is subjected to a pressure of many atmospheres, the temperatures would be raised in the case of those which contract on freezing, and lowered in the case of substances like water, which expand.

A little consideration shows that this effect is almost to be expected, because pressure tends to reduce volume, and this should facilitate a process involving shrinkage, while it might equally be expected to make change of state more difficult when the substance expands in the process. Thus when ice melts it contracts, and pressure makes the process easier, resulting in its occurring below the normal temperature. On the other hand, rocks expand during fusion, so that pressure

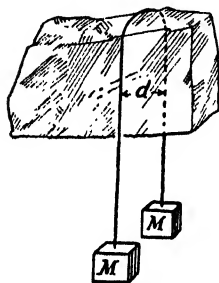


Fig. 13.

resists the process and raises their melting point. We may then make the general statement that *pressure favors the state of smallest specific volume*, and consequently shifts the melting point in the direction of larger volume.

Ice at 0° may be melted by the pressure exerted by a fine wire passing over it with weights at its ends, as shown in Fig. 13. The ice immediately under the wire melts and, as water, passes around the wire to be frozen again above it under reduced pressure. Thus the wire passes

slowly through the block of ice, leaving it as solid as it was before.

As will be explained later, the melting process involves an absorption of heat tending to lower the temperature of the water formed under pressure, but above the wire, where it freezes again, this heat is released and flows back across the wire to help in melting some more water. Consequently, a wire of high thermal conductivity like copper cuts its way through the ice faster than one of lower conductivity, provided both have the same diameter and exert the same pressure.

The pressure under the wire may be roughly calculated if we know the diameter, d , of the semicircle in which it soon forms itself, and its own sectional diameter. The product of these two quantities gives the effective area upon which the force $2Mg$ acts. If M is 1 kg, d 10 cm, and the sectional diameter 0.05 cm, the pressure is given by

$$\frac{2000 \times 980}{10 \times 0.05} = 3.92 \times 10^6 \text{ dynes/cm}^2,$$

or nearly four atmospheres. As one atmosphere lowers the melting point 0.0075 degree, the result is a lowering of about 0.03 degree.

The motion of glaciers under the tremendous pressure to which their lower portions are subjected is at least partly due to this phenomenon, while the liquefaction of rocks in forming lava under reduced pressure represents the reverse case. One theory of volcanic eruptions accounts for lava flows by supposing the rocks at a great depth to be at a temperature high enough to melt them under ordinary conditions, but that the pressure of the strata above keeps them in the solid state. If this is relieved by a slipping, or other readjustment of the earth's crust, the rocks melt, and their increased volume in the liquid state finds an outlet by way of a volcano.

Under enormous pressures water freezes into several different sorts of ice, some of them denser than water. Ice of such character is harder, instead of easier, to melt under pressure, so that its melting point is raised. Professor P. W. Bridgman of Harvard University, who has examined the effects of enormous pressures on a great variety of substances, finds that under a pressure of nearly twenty thousand atmospheres, the abnormal ice just mentioned must be heated to 76°C to melt it.

220. Heat of fusion. If a jar containing ice and water is placed on a stove, the ice begins to melt and this process continues until the mixture is wholly liquid. If the contents of the jar are well stirred while the melting continues, a thermometer placed in the water registers exactly zero, and does not begin to rise until the ice has disappeared. Evidently the heat which was supplied has done something quite different from its usual effect of raising the temperature of the body being heated. The energy of the heat thus supplied to change a body's state from solid to liquid has been utilized in changing the molecular structure of the body, by giving the molecules the greater mobility which they exhibit in the liquid state. The *amount of heat necessary to change one gram of a substance from the solid to the liquid state at constant temperature* is designated by the letter L and has long been known as the *latent heat* of fusion. But it is now considered better to call it simply the *heat of fusion* without using the word "latent."

In the reverse process, L calories must be withdrawn from the liquid at the freezing point to transform it into the solid state at the same temperature. Thus, since heat flows *into* a body when it melts, fusion tends to cool surrounding objects. The stove on which the ice is melted is kept cooler during the process than it would have been otherwise. Similarly freezing is a heating process, because heat flows *out* of a body while freezing is going on, and tends to maintain its

surroundings at the freezing point when they might otherwise become much cooler.

We can now understand why the water formed from ice melted by pressure tends to be cooled. The heat of fusion must be supplied from somewhere, and if no external heat is available, the liquid product of the fusion itself supplies the needed calories, and is cooled in consequence.

As an example of the reverse process, a tank of water in an unheated cellar in winter maintains the temperature at the freezing point until it is frozen solid, because it is constantly giving off heat of fusion to its surroundings.

221. Value of heat of fusion. It requires 79.6 calories to change one gram of ice at 0°C to water at the same temperature, though in most calculations 80 calories is sufficiently accurate. The values of L in calories per gram for some other substances are given in the following table:

Elements	$t (^{\circ}\text{C})$	L	Compounds	$t (^{\circ}\text{C})$	L
Aluminum.	657	77	Ammonia.	-75	108
Lead.	327	6	Sulphuric Acid. . . .	10.3	24
Mercury.	-39	3	Benzene.	5.4	30
Silver.	960	21	Acetic Acid.	16.7	43
Tin.	232	14	Glycerine (pure). . .	18	48

Heats of fusion are frequently measured by the method of mixtures. In this case, in addition to the heat of the body gained or lost in its solid state, a similar item for its liquid state must be introduced, and a third involving the heat of fusion. Thus if Σsm represents the water equivalent of the calorimeter and its contents, m_x the mass of the body under examination, s_s and s_l its specific heats as solid and liquid, and L its heat of fusion, then

$$\Sigma(sm) \times (t_3 - t_2) = s_l m_x (t_1 - t_m) + L m_x + s_s m_x (t_m - t_3),$$

where t_m is the temperature of the melting point. This assumes that the body x was at first liquid at a higher temperature t_1 than the calorimeter. Then it would have to be poured into a container immersed in an oil or water bath whose original temperature was t_2 , and lower than its final temperature t_3 .

However, the measurement may be made the other way; for example, by putting a piece of ice into a bath of water, and noting the final temperature. If the ice was originally at t_1° below zero, its

absorption of heat up to 0° must be allowed for, and the equation becomes

$$\Sigma(sm) \times (t_2 - t_3) = s_i m_i (0 - (-t_1)) + L m_i + s_w m_i (t_3 - 0),$$

where the subscripts i and w indicate the state, and where the final temperature of the calorimeter, t_3 , is lower than the initial temperature t_2 . Either of these equations is easily solved for L , so that it may be calculated when all the other quantities are known.

222. Supercooling. Liquids which crystallize on freezing may be cooled below their normal freezing points, provided they are kept very still during the process. Fahrenheit, who first discovered this fact, sealed water in a glass bulb before cooling it, and Gay-Lussac cooled water to -12°C without the formation of ice, by covering it with a layer of oil. A sudden disturbance of the supercooled liquid, however, or the introduction of a crystal of the solid substance, results in sudden freezing.

When a supercooled liquid freezes, the heat of fusion warms the whole mass till it reaches the normal melting point. Then the process of solidification ceases, with both solid and liquid states in equilibrium with each other. Or if the initial temperature is sufficiently low, the entire mass may be frozen.

Since pressure tends to produce solidification of most substances, certain supercooled liquids may be frozen with a sudden evolution of heat by subjecting them to an increased pressure. A solid like sodium hyposulphite, which normally melts at 49.5°C , may be supercooled to room temperature and kept indefinitely in the liquid state, but if compressed, the mass solidifies with an evolution of heat, and the temperature rises once more to the melting point. Thus part of the heat which was supplied when the compound was first melted remains in the form of unstable potential energy ready to be released as the kinetic energy of heat, when desired.

223. Vaporization. The transformation of a liquid into the gaseous state is called **vaporization**, and the product is called a **vapor**. This is because gases near the point of liquefaction deviate so far from the laws of Boyle and Charles that they are given a different name to distinguish them from the same substance at a much higher temperature. If this process occurs only at the free surface of the liquid, it is called **evaporation**, but if it takes place all through it, the liquid is said to *boil*. The reverse of these processes, when the vapor becomes liquid, is called **condensation**.

It is also possible for a solid to pass directly into the gaseous state

without first becoming liquid. It is then said to *sublime*, and the process is called **sublimation**.

224. Evaporation. The molecules on the free surface of a liquid have more latitude of motion than those in its interior, and are constantly escaping into the space above it, away from the forces of cohesion which hold them back. This tendency increases with the temperature, which indicates a gain of kinetic energy and a higher probability of escape. The pressure exerted by these escaping molecules is exactly like that of a gas within a container, and is known as **vapor pressure**. If it is in excess of the pressure of the same vapor in the space immediately above the liquid, the process of evaporation continues. But when these pressures are equal, as many molecules return per second to the liquid as are leaving it, and evaporation ceases. This state of things is known as statistical equilibrium, as in a town when the birth and death rates are equal. If this is true, the total population remains constant, though the individuals change.

If the space above the liquid has a greater vapor pressure than the liquid itself, more molecules enter than leave the surface, and condensation takes place.

The exact conditions which determine these processes will be explained later, but enough has been said to account for the general mechanism of this kind of change of state.

225. Boiling. In order to boil, the liquid must form vapor all through its mass. But vapor tends to occupy a much greater volume than the same mass of liquid. This means that the vapor has to push the liquid aside and form bubbles. But bubbles within the liquid are subjected to the atmospheric pressure acting on its surface plus the pressure due to the liquid itself at the depth where they are formed. Therefore if a bubble is to form, the vapor pressure (or vapor "tension") within it must exceed the total pressure at that point. As soon as this occurs, the vapor is disengaged all through the liquid, and the resulting bubbles rise and burst at the surface, so that vaporization is much more rapid than is possible when it occurs only at the free surface.

226. The boiling point. The temperature at which the vapor pressure of the liquid just equals the pressure on its surface, is known as the **boiling point**. This differs widely with different liquids, and of course varies with the pressure. If the pressure over the liquid is lowered by an air pump, or by ascending to altitudes where the barometer stands lower than at sea level, the liquid boils at a lower temperature than its normal boiling point under a standard

atmosphere. On the other hand, liquids in a boiler under pressure boil at a higher temperature than their normal boiling points.

The boiling points under normal atmospheric pressure of a number of liquids whose values are known with precision, are given in the accompanying table. These are obtained from the temperature of the vapor just over the boiling liquid, because the liquid itself at points below the surface must be at a slightly higher temperature in order to form bubbles under a pressure a little greater than on its free surface.

Substance	Boiling Point
Oxygen.....	-183° C
Ammonia.....	-38°5
Ether.....	34°6
Alcohol.....	78°3
Glycerine.....	290°
Mercury.....	356°7
Zinc.....	907°

227. Superheating. When a liquid in a glass jar is brought gradually to a boil, it is often noticeable that the bubbles seem to originate from certain particular points of the container, and if a few grains of sand are thrown in they are seen to act as nuclei for the formation of the vapor, and so greatly facilitate the process of boiling. These centers are either minute air bubbles adhering to the glass, or specks of dust, or other impurities, and it can be shown that without their aid the formation of a bubble in the liquid is much more difficult.

Now suppose a liquid has been freed from impurities and previously boiled to get rid of all dissolved air; then after cooling, it may be heated in a clean and smooth receptacle above the boiling point without ebullition taking place. In this way water may be *superheated* to 137° C before boiling begins. But then it takes place with the explosive violence which is called "bumping."

Even without especial precautions water may fail to boil at exactly 100°, provided it has been boiled before, and when this delayed boiling finally takes place at a few degrees above 100°, the resulting bumping may throw some of the water out of the boiler. A spoon is sometimes put into a jar of water to prevent this effect, but the introduction of some porous substance like charcoal is the surest preventive of superheating.

228. Heat of vaporization. The change of state from liquid to vapor requires a large amount of energy which must be supplied in the form of heat, just as in the case of fusion. But vaporization is a more radical change than fusion, not only in giving the molecules a much greater freedom of motion, but also in separating them much more widely. Therefore, as we should expect, the heat of vaporization is in general much greater than that of fusion. The heat of

vaporization, designated by L , is the number of calories required to evaporate one gram of a liquid at constant temperature, and it varies with the temperature at which the vaporization takes place.

According to Henning, the heat of vaporization of water is given approximately by $L = 538.86 + 0.5994(100 - t)$, for a range between 30° and 100° C. Therefore if $t = 100^\circ$, the heat required to evaporate one gram of water at that temperature is 538.86 gram calories, though 540 is a more recent value. If Henning's formula is assumed to hold good at 0° , the heat required at that temperature is about 599 calories. It is actually about 596.

The energy represented by L , as has been stated, gives the molecules greater freedom, and increases the volume they occupy when the liquid becomes a vapor. In the case of water boiling at 100° , about one thirteenth of the heat of vaporization is concerned in the work required to overcome the external pressure, while the remainder increases the *intrinsic energy* of the substance in giving the molecules greater freedom of movement. The former item disappears under zero pressure, which would seem to mean that L should then be less than before. This is contrary to fact, however, for under zero pressure, water boils very close to zero degrees, and $L = 596$ calories per gram instead of 540 when under atmospheric pressure. This apparent contradiction is explained as follows: The gain in eliminating the work done against the atmosphere is more than offset by the great increase in the change of intrinsic energy when cold water is vaporized, above that involved in vaporizing hot water.

The following table gives in calories per gram the heats of vaporization, under atmospheric pressure, of some of the more familiar liquids, as well as of substances normally in the gaseous or solid state.

Substance	Temperature ($^\circ$ C)	Heat of Vaporization (Calories per gram)
Water	100	540
Ammonia	-38.5	341
Alcohol (ethyl)	78.3	208
Ether	34.6	91
Hydrogen	-253	108
Air (liquid)	-196 (N_2)	51
Mercury	356.7	68
Sulphur	316	362

229. Condensation. If heat must be added to vaporize a liquid, it must be withdrawn to liquefy a vapor. The amount required per

gram is the same either way for bodies with fixed boiling points, so that when one gram of steam at 100° condenses to water at 100° , it gives out 540 calories of heat. This principle is made use of in steam heating systems where the steam condenses in the radiator and gives up large amounts of heat in so doing.

230. Cooling by evaporation. When a liquid evaporates, the energy required by the transformation is withdrawn from its surroundings, which are cooled in consequence, unless heat is supplied at the same time. The cold sensation of the skin when wet, especially in a current of air, is due to evaporation of the water, while a volatile liquid like ether held in the hand may feel painfully cold because of its rapid vaporization.

In hot climates water is kept in porous earthenware jars, so that it "sweats" through them and evaporates from the outer surface, thus cooling the entire contents, while perspiration is nature's method of lowering the temperature of the body, even below that of the surrounding air.

231. Joly's steam calorimeter. One of the most sensitive and accurate calorimeters was designed by Dr. Joly in 1886. It makes

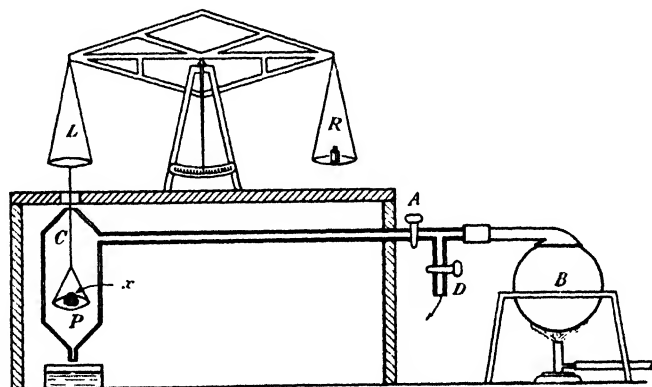


Fig. 14.

use of the condensation of steam and consequent liberation of heat of vaporization to determine specific heats. The essential principle is that a body of mass M at a temperature t lower than 100° C, when surrounded by saturated steam at the boiling point, causes some of the steam to condense over its surface, thereby increasing its apparent weight. The amount m of water condensed multiplied by the heat of vaporization L gives the heat lost by the steam. This is equal to

the heat gained by the body in having its temperature raised to 100° . Therefore $mL = Ms(100 - t)$, where s is the specific heat required.

The main features of the apparatus are shown in Fig. 14. The substance whose specific heat is to be measured is placed in a pan P and steam from the boiler B is admitted by the valve A , with D closed, into the calorimeter C . The increased weight is found by adding weights to the pan R first with P empty and then with the object added. The difference in the two weights of water condensed is due to the object.

232. Sublimation. Certain solids under ordinary conditions pass directly into the gaseous state, that is, **sublime**, or may be made to do so by heating them. Camphor is a familiar substance having this property, for it remains perfectly dry while giving off the vapor used in protecting woolen fabrics from moths. Its mass steadily diminishes during this process, but if kept in a tightly closed vessel, the vapor may be condensed to form solid camphor again, just as ordinary vapor condenses to a liquid on being cooled.

Iodine and mercuric chloride (corrosive sublimate) are other substances which evaporate directly from the solid state when heated under ordinary atmospheric conditions. Even ice and snow evaporate at temperatures below freezing, and solid carbonic acid (now marketed as "dry ice"), when exposed to the atmosphere, passes rapidly into the gaseous state without liquefying.

We shall see later that many other substances may be made to sublime under unusual conditions, while on the contrary, those named above may be liquefied by heat, provided the pressure exceeds a certain minimum value higher than that of normal atmosphere.

SUPPLEMENTARY READING

C. H. Draper, *Heat* (Chap. 8), Blackie & Son, London, 1911.

P. W. Bridgman, *The Physics of High Pressure*, G. Bell & Sons, London, 1931.

PROBLEMS

1. A lump of ice at 0°C is dropped into 150 cm^3 of water at 80°C . When it is all melted the volume is 196 cm^3 . If the effect of the calorimeter is negligible, what is the final temperature? *Ans.* 42.4°C .

2. A piece of ice at -20°C and weighing 57 g is dropped into 220 cm^3 of water at 75°C . The copper calorimeter containing the water weighs 240 g , and the final temperature is 43.5°C . What is the specific heat of the ice? *Ans.* 0.50 calories per gram.

3. Show that the heat of fusion of water is $144\text{ B.t.u. per pound}$.

4. A shower of rain precipitates 6 cm on a frozen lake. The temperature of the water is 12°C and that of the ice 0°C . What thickness of ice is melted? (The density of ice is 0.917 g/cm^3 .) *Ans.* 9.8 mm.

5. An unheated shed in winter is kept at 0°C by the gradual freezing of a tank containing 1000 kg of water. If the shed loses heat through its walls at an average rate of 5000 calories per minute, how long can it be kept at 0° ? *Ans.* 11.1 days.

6. Two weights of 2.5 kg each are suspended by a wire passing over a block of ice, as in Fig. 13. The wire has a sectional diameter of 0.02 cm and forms a circular arc 8 cm in diameter. How much is the freezing point lowered under the wire? *Ans.* 0.23° , nearly.

7. How much heat is required to raise 240 g of water at atmospheric pressure from 20°C , and convert it into steam at 100°C ? *Ans.* 148,800 calories.

8. Calculate the latent heat of vaporization of water in B.t.u. per pound from its value in calories per gram. *Ans.* 972 B.t.u. per pound.

9. How much heat is required to convert one pound of ice at 0°F to steam at 212°F ? (The specific heat of ice is 0.5.) *Ans.* 1312 B.t.u.

10. Steam at 100°C is passed into 344 g of water contained in a copper calorimeter at 20°C until its weight has increased 36 g, and the temperature is 76.5°C . The calorimeter weighs 120 g. What is the observed heat of vaporization? *Ans.* 535 calories per gram.

11. Six liters of water at 30°C stand in a porous jar and "sweat" through the jar into air at the same temperature. Calculate the latent heat at 30° from Henning's formula (Article 228). Calculate the temperature of the water, after 50 cm^3 have evaporated, neglecting the effect of the jar itself and assuming L constant. *Ans.* 580.8 calories per gram; 25.2°C .

12. How many calories are required to evaporate 60 g of water under a pressure of 9.2 mm of mercury? (Use table of Article 235.) *Ans.* 35,604 calories.

13. A boiler containing 60 l of water at 120°C bursts. How much of the water is vaporized? (Take 532 as average value of L .) *Ans.* 2.26 l.

14. The specific heat of air at constant pressure is 0.2417 at 20°C and its density is 0.0012 at 20°C and 76 cm of mercury. How much is the air of a room, $3 \times 4 \times 6$ m in size, warmed by the condensation of 600 g of steam to water at 100°C in a radiator, if losses are negligible? *Ans.* 15.5° .

15. Using Henning's formula, calculate the amount of heat required to bring a gram of water from 80°C to 100°C and vaporize it completely. Then compute the specific heat of the steam by setting this result equal to the heat required to vaporize the water at 80° and bring the resulting steam to 100° at atmospheric pressure. *Ans.* 558.86 calories; 0.4 calorie per gram. (NOTE: This is about halfway between the accepted constant pressure and constant volume values of water vapor.)

CHAPTER 18

Vapors and Gases

233. Saturated vapors. When a liquid evaporates into a confined space, the process continues until the vapor pressure within that space becomes equal to the pressure of evaporation from the free surface of the liquid. The process then ceases, the vapor is said to be saturated, and its pressure is added to that of any other vapor or gas that may be there already, as will be explained more fully in Article 236. If now the liquid is heated, its vapor pressure rises, more vapor escapes, and the space once more becomes saturated at the new and higher temperature. Thus it is seen that saturation is a relative term, and that the higher the temperature, the more vapor is required to produce this condition.

Let us now suppose that the temperature is kept constant, and that the confined space above the liquid is occupied by the vapor alone.

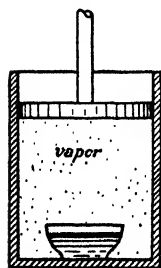


Fig. 15.

Then imagine that the volume can be either increased or decreased, as by a piston sliding in a cylinder, as indicated in Fig. 15. If the piston is raised and the volume increased, the tendency to lower the pressure by expansion is quickly overcome by renewed evaporation from the free surface of the liquid, while if the piston is lowered, the tendency to raise the pressure results in condensation either on the free surface of the liquid, or on the walls of the enclosure. In either case the pressure remains the same, and is thus

shown to be independent of the volume, though it does depend upon the temperature which was supposed constant during the process.

If there is sufficient liquid in the enclosure, so that it cannot completely evaporate, neither changes of temperature nor volume can alter the final result of filling the space with saturated vapor, which is then in *statistical* equilibrium with its liquid. We may define saturated vapor as a *gaseous body in equilibrium with its liquid when confined within a limited space.*

234. Measurement of temperature-pressure relations. Since the pressure of a saturated vapor depends only upon its temperature, it

is most important to know how it varies with the temperature. This is especially true of steam because of its importance in steam engines.

The classical determination of these relations was made by Regnault, whose apparatus is sketched in Fig. 16. The copper boiler *A*, partly filled with water, has four tubes closed at their lower ends, into which are inserted four thermometers which are thus kept from direct

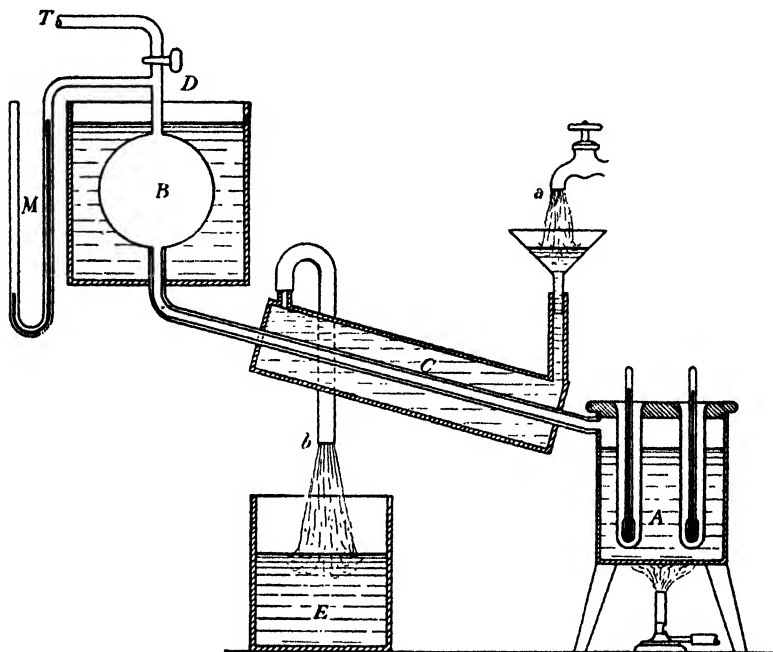


Fig. 16.

contact with the liquid, but whose average reading gives its temperature. From the space above the water, a brass tube passes to the globe *B*, which is surrounded by a water bath to keep its temperature constant. A water jacket *C*, through which runs a constant stream of cold water, serves to condense the steam as it passes up the inner tube. The manometer *M* shows the pressure in *B* and *A*, while a tube *T* equipped with a valve *D* is connected to an air pump so that the pressure within the apparatus may be lowered at will.

Suppose the pressure has been reduced to 17.5 mm of mercury; then if the water is at 20° C, it begins to boil, while the resulting steam is condensed by *C*, and runs back as water into the boiler. This could not continue unless heat were constantly supplied to the boiling

water, because the removal of the heat of vaporization tends to cool it below 20° , and the condensed steam returning to it may be cooler still. If, however, a small flame below the boiler is kept burning, the water will boil at 20° indefinitely. The heat thus supplied is taken up by the formation of steam and then given over to the cooling water in the condenser, when the steam is condensed. Thus the water escaping at b is warmer than at a where it enters the jacket, so that all the calories supplied by the flame ultimately appear in the water collected in the tank E , or are radiated from the boiler and tubes.

To obtain the vapor pressure corresponding to a temperature above the initial one, air may be admitted through the valve D . This stops the boiling at once, but it is resumed as soon as the vapor tension of the water in A again equals that of the air above it. New readings of the manometer and thermometers are then taken, and so a succession of points on the temperature-pressure curve may be obtained up to a full atmosphere, or even higher, if the apparatus is designed to withstand high internal pressure, and if an air compressor is substituted for the vacuum pump.

235. The pressure-temperature curve. The results of the preceding experiment are shown in the curve of Fig. 17. It does not begin quite at the origin because both ice and water exert a small vapor pressure of 4.579 mm at 0° . After that the pressure rises at a steadily increasing rate per degree, and at 100° , when the pressure is one atmosphere, it is very steep, so that small temperature changes make large pressure changes. The exact values of the pressure for every ten degrees are given below. Steam tables giving these values

$t (^{\circ}\text{C})$	p in mm	$t (^{\circ}\text{C})$	p in mm
0	4.579	60	149.2
10	9.205	70	233.5
20	17.51	80	355.1
30	31.71	90	525.8
40	55.13	100	760.0
50	92.30	110	1074.5

for every degree are available, and are of the utmost practical value to steam engineers. The relations shown by the curve and tables may be thought of as representing both the boiling points of the liquid under the atmospheric pressures assigned to them, and also the pressure of saturated vapor at the given temperatures. The first of these two aspects follows from the fact that when a liquid boils, its vapor pressure equals that exerted by the atmosphere above it. This is the case whether the atmosphere is of saturated vapor only, as in an engine boiler, or whether it consists of air and vapor combined, as when a liquid boils in an open pot.

The second aspect applies to saturated vapor pressure, either with or without the presence of an excess of the liquid, and the latter need not boil to give off the vapor at a pressure corresponding to its temperature.

Thus saturated steam at 50° has a vapor pressure of 92.3 mm. If the atmospheric pressure is greater than this, water present does not boil, but it evaporates with a pressure of 92.3 mm until the space above it is saturated. If the atmospheric pressure is lowered to this value, the only change is that the water now boils and gives off vapor more rapidly in consequence.

In Fig. 17 the space above the curve is a liquid region where the pressures are too high for boiling at their corresponding temperatures. If the pressure and temperature of a liquid are given by the co-

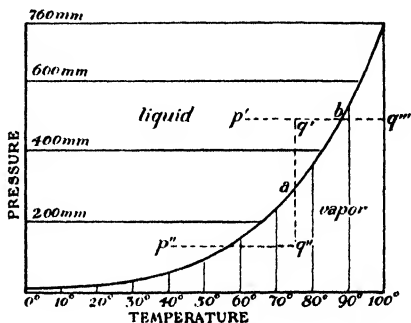


Fig. 17.

ordinates of the point q' , and if the vapor above it is saturated, it remains permanently in the liquid state. In the space below the curve, only vapor can exist in equilibrium, for the pressure at q'' with the same temperature as at q' is lower than at the boiling point a , and liquid then would boil violently and be all converted into vapor. Or we might say that q' represents the liquid state, because the temperature is too low for boiling under the existing pressure, while q'' is in the vapor region because the pressure has been lowered below a , where boiling begins without change of temperature. Also, q''' is in the vapor region because the temperature has been raised above the boiling point b at the same pressure.

Saturated vapor, if raised to a temperature higher than that which corresponds to its pressure, as indicated by q''' , is said to be superheated. This is the usual condition of the moisture in the air about us, for its pressure is usually less than would be required for saturation. On the other hand, a lowering of the temperature below that required for saturation results in supersaturation, with resulting condensation in fog, dew, and so forth.

236. Mixtures of vapors and gases. The pressure just referred to is only that of the vapor present in the air, and not that of the atmosphere as a whole. Variations of the former would affect con-

densation or evaporation, while the *total* pressure influences the boiling point only. This statement involves the important principle that the vapor pressure may be regarded as distinct from the pressure of the air into which the liquid evaporates. Dalton (1766–1844) investigated the pressure of mixtures of this sort, and formulated two laws, the first of which states that *the partial pressures of one or more vapors mixed with a gas are the same as if each filled the space alone*; and the second law states that *the total pressure is the sum of all the partial pressures of gases and vapors present in an enclosure*.

These laws seem somewhat paradoxical until we recall the fact that the spaces between the molecules of a gas are so great compared to the volumes of the molecules themselves at ordinary pressures, that many more may be introduced without appreciably interfering with the motions of those already there. Then if there is no interference, each acts as if the other were not present, while the total pressure would be exactly the sum of the pressures exerted by each when present alone.

It makes, however, a decided difference whether a liquid evaporates into a vacuum or into a space already occupied. It fills the vacuum almost instantaneously while it boils with explosive violence. But when a gas is already there, it takes some time to saturate the space, and the liquid may or may not boil according to whether the pressure on the free surface is less or greater than its vapor tension. The ultimate vapor pressure is the same in either case, and after the space has become saturated, the final pressure is that of the air p_a plus the vapor pressure p_v , and the total atmospheric pressure $P = p_a + p_v$. If other gases or vapors are introduced, their partial pressures must be added, though each takes a little longer than its predecessor to reach saturation.

It is now more easily understood why a glass of water at room temperature usually evaporates. If the air is not already saturated with water vapor, its pressure p'_v is less than the pressure p_v exerted by the vapor at the surface of the liquid. Therefore, though the presence of the air prevents boiling and retards the process of saturating the room, the water continues evaporating until the partial pressure of the water vapor in the air about it equals its own, which in this case means saturation, as both water and vapor are assumed to be at the same temperature.

237. Condensation by expansion. If the bell jar of an air pump becomes saturated with water vapor at atmospheric pressure from a dish of water placed inside, a few strokes of the pump cause a cloud

to form which disappears again when the air is allowed to rush in to fill the partial vacuum. This is because the air is cooled by the sudden expansion, as we have seen in Article 201. Although the pressure is also reduced, the path followed by the vapor, indicated by cf in Fig. 18, brings it within the liquid region, and it condenses on minute dust particles in the air. When the original pressure is restored, the heat of compression warms the droplets, following the curve from f to c where they evaporate.

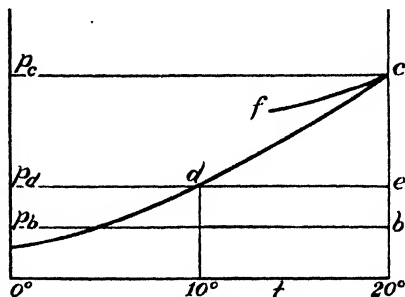


Fig. 18.

The cumulus clouds which form on a hot afternoon in summer are produced in this manner. Columns of moist heated air rise to a height of several miles, where the expansion under reduced pressure cools them, and condensation results. But if the air is dust free and has no other nuclei on which moisture may deposit, the clouds form with difficulty, and then only when the change in pressure and consequent cooling are relatively great.

238. Humidity. If the air in a room is at 20°C , the space it occupies is undersaturated when it contains vapor at the pressure corresponding to the point e shown in Fig. 18. But at 10° the same vapor at the same pressure would saturate it, because d lies on the saturated vapor curve. This means that the same amount of vapor which saturates the space at a lower temperature fails to do so at a higher one, and we cannot tell how near a given vapor pressure comes to saturating a space until we know the temperature.

In order to form a picture of the degree of saturation of the space, the idea of *relative humidity* has been introduced into the science of the weather known as *meteorology*. This is defined as *the ratio of the amount of water vapor actually present in the air, to the amount necessary to saturate it at the same temperature*. As these amounts or masses contained in a given volume vary almost exactly as their pressures, relative humidity may also be defined in terms of pressure, so that in Fig. 18, if e represents the condition of the vapor actually present in the air at 20° , the relative humidity is p_e/p_c , for p_c is the pressure of saturated vapor at 20° . If the air were now cooled to 10° , its relative humidity would be $p_e/p_d = 1$, or 100 per cent, as compared to about 40 per cent in the previous case.

There are three general methods for measuring relative humidity. We may pass a known volume of air through drying tubes and weigh the total moisture removed from it, and compare this with the amount known to exist in the same volume when saturated. Or we may cool the air at constant pressure, following a course like *ed* in Fig. 18, and find the temperature when the air becomes saturated. This temperature is called the **dew point**, because when *d* is reached, moisture begins to form on the surfaces in contact with the chilled air. Then the ratio of the dew-point vapor pressure to the saturation pressure at the temperature of the unchilled air gives the desired result. The third method is to compare the readings of two thermometers, one of which is dry, while the bulb of the other is surrounded by a cloth or wick moistened with water whose evaporation lowers its temperature. The lower the relative humidity, the cooler the wet bulb becomes, and the desired value may be obtained from tables based on dew-point measurements, which give the relative humidity corresponding to the "wet and dry bulb" readings. Such instruments are known as **hygrometers**, and the science of measuring humidity is called **hygrometry**.

The comfort or discomfort we experience as a result of climatic conditions depends quite as much upon relative humidity as upon temperature. We suffer far more on a "muggy" day in summer when the thermometer reads 85° F than on a dry day at 95° F because the higher relative humidity at the lower temperature retards evaporation of moisture from the skin, and so retards the resulting lowering of body temperature. We are also more chilled by a damp day in winter than by "dry cold" at the same or even a lower temperature, probably because moist skin is more sensitive to either heat or cold.

On the other hand, a certain amount of moisture in the air is highly desirable. Without it evaporation from the pores is too rapid, the skin dries and cracks, and our general health is impaired. This state of things is common indoors in winter. Suppose the air outside has a relative humidity of 90 per cent at a temperature well below freezing. This is a high humidity, but when the same air is heated to room temperature, 20° C (or 68° F), its relative humidity may be as low as 30 per cent, although no moisture has been removed from it. It is not necessary to overheat a house to obtain this result, which depends rather upon the low temperature out of doors. Air as dry as 30 per cent relative humidity tends to take up moisture wherever it may be found, or rather, moist objects evaporate into it with great rapidity. So our furniture dries and cracks, and our hands

get chapped even if we never leave the warmth of the house. These conditions may be corrected by a spray of minute droplets of water injected into the hot air stream which warms a house heated by hot air, or the spray may be produced in individual rooms. Such devices come under the head of "air conditioning" and are increasingly popular.

SUPPLEMENTARY READING

C. H. Draper, *Heat* (Chap. 9), Blackie & Son, London, 1911.

Saha and Srivastava, *A Textbook of Heat* (Chap. 5), Indian Press, Allahabad, 1931.

T. A. Blair, *Weather Elements* (Chap. 3), Prentice-Hall, 1937.

PROBLEMS

1. The volume swept out in one stroke of a certain steam pump is 300 l. The temperature of the steam (saturated) is 180°C . At this temperature, its pressure as found in the steam tables is 10.216 kg/cm^2 . Calculate the work done in one stroke of the piston, assuming constant pressure. *Ans.* 3.004×10^5 joules.

2. The density of saturated steam at 180°C is 5.15 grams per liter. Calculate the heat required to produce the steam needed for one stroke of the pump in Problem 1, taking the latent heat of vaporization as 480.6 calories per gram. *Ans.* 742,527 calories. (NOTE: As will be seen farther on, this represents 31.07×10^5 joules, or about ten times as much energy as the amount realized in the stroke of the piston.)

3. A dew-point hygrometer indicates condensation at 10°C . The surrounding air is at 30°C . Using the steam table in Article 235, calculate the relative humidity. *Ans.* 29 per cent.

4. How cold must a glass of water be to form moisture on its surface when the relative humidity is 57.5 per cent and the temperature is 104°F ? *Ans.* 86°F .

CHAPTER 19

Relations Between the States

239. The triple point. Substances having definite freezing and boiling points may exist in all three states of solid, liquid, and vapor. These states are sharply defined, and on the pressure-temperature

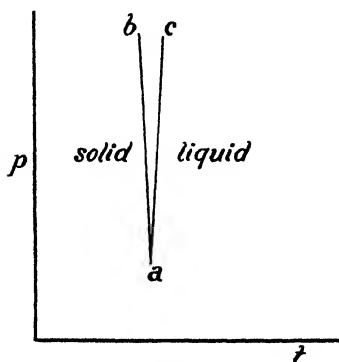


Fig. 19.

diagram their bounding conditions are indicated by lines such as the saturated vapor curve already discussed. A similar line marks the boundary between solid and liquid, such as either *ab* or *ac* in Fig. 19. The line *ab* belongs to those bodies which, like water, expand on freezing and so have their melting point lowered by pressure, while *ac* represents the more usual case where pressure raises the melting point.

A line similar to the saturated vapor line defines the boundary between solid and vapor, as shown in Fig. 20. Here the pressures are extremely low, even in the case of such volatile substances as camphor. But the curve is concave upward, showing a tendency to reach much higher pressures if the solid state could be maintained at temperatures higher than those usually employed.

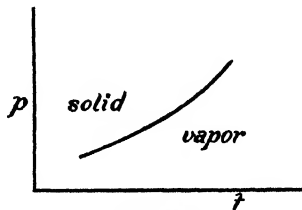


Fig. 20.

These three curves—liquid-vapor, solid-liquid, and solid-vapor—meet at a common **triple point** where all three may exist together in equilibrium at a definite temperature and pressure. In Fig. 21 the three curves which separate ice, water, and steam are shown for a small range around the freezing point, near which the triple point is located. The sublimation curve and curve of boiling form a nearly continuous smooth curve, though not quite, for the former is steeper

near P than the latter. The fusion line really slopes very slightly to the left, like ab in Fig. 19, but this is too slight to be seen in the diagram. The triple point P is at a pressure of 4.6 mm and a temperature of 0.0076° above zero centigrade.

The curves of Fig. 22 are drawn without reference to any scale, in order to bring out certain interesting features of the diagram. The line ab shows what happens if ice is heated under constant pressure higher than 4.6 mm. It turns to a liquid when the fusion line is crossed, and at a little higher temperature, t_1 , the liquid boils and becomes vapor which is in the superheated state at b . A similar process at a pressure below 4.6 mm would result in vaporizing the ice without liquefaction when it crosses the sublimation curve at $-t_2$. Lowering the pressure over a liquid from c at constant temperature t_3 causes it to boil, when it crosses the curve of boiling, and to become a superheated vapor at d . But a similar process starting at a very high pressure indicated by c' , and a temperature just below that of the triple point, would cause water first to freeze and then become vapor at d' .

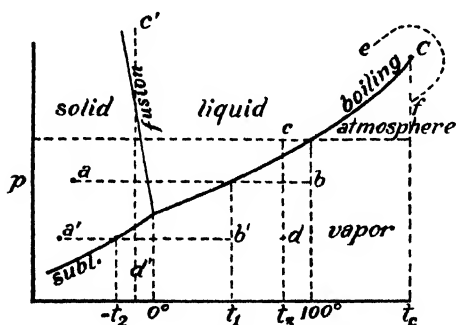


Fig. 22.

to carry a liquid over into the vapor region by the path cf . The liquid becomes increasingly gaseous until it has acquired all the properties of a vapor, without any abrupt change in its state. No such end of the other two curves has ever been found.

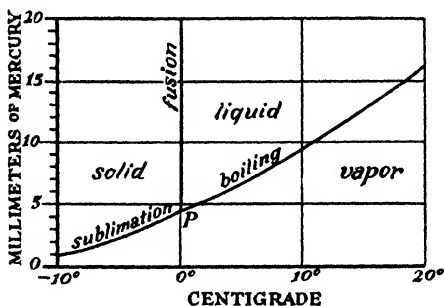


Fig. 21.

then become vapor at d' .

Owing to the fact that the latent heat of vaporization diminishes with rising temperature, and finally vanishes at 365°C in the case of water, boiling is impossible at temperatures higher than this. Therefore the curve of boiling comes to a definite end, as indicated in the diagram, and it is possible

240. Freezing by boiling. A striking experiment may be performed which brings water to the triple point, where it boils and freezes simultaneously. A watch glass containing water is supported by a

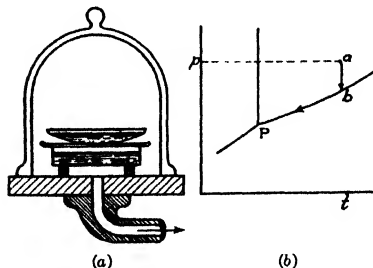


Fig. 23.

light wire "spider" over a dish containing concentrated sulphuric acid to absorb water vapor, as shown in Fig. 23 (a). This arrangement is covered by the bell jar of an air pump, and the pressure over the water progressively reduced. If the initial conditions are represented by *a* (Fig. 23 (b)) the effect of working the pump quickly lowers the pressure to *b*, when boiling begins. This lowers the temperature of the water, owing to the loss of the heat of vaporization, and it now follows the curve of boiling from *b* to *P*, with steadily decreasing temperature and pressure. At *P* the ice phase is reached and freezing begins, with bubbles bursting through the thin layer of ice as it forms. If the sulphuric acid were not used, the vapor formed from the boiling water would continually tend to raise the pressure again, and the requisite value of 4.6 mm could not be obtained until all the water had been evaporated.

241. The critical point. The temperature t_c which marks the end of the curve of boiling is known as the **critical temperature** of the substance, and the point *C* (Fig. 22) is the **critical point**. At a lower temperature than the critical, an increase of pressure upon the vapor causes it to pass from the vapor state, as at *d*, to the liquid state at some point *c*, after crossing the curve. But no amount of pressure can liquefy a gas or vapor if the temperature is maintained above t_c .

This important property of gaseous bodies is better shown on the pressure-volume diagram, Fig. 24. The curves are all isothermals of

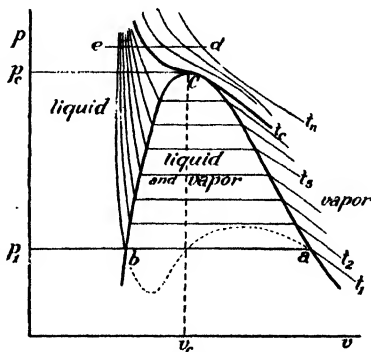


Fig. 24.

ascending temperatures from t_1 to t_n , and t_c is the critical temperature. If a vapor at a temperature t_1 below the critical value is compressed, its volume decreases as the pressure rises, until at a point *a* it becomes

saturated and begins to liquefy. Since saturated vapor has only one pressure at a given temperature, the pressure p_1 now remains constant and the volume diminishes until the point b is reached. The substance is now wholly liquid, and further increase of pressure results in very slight changes in volume as indicated by the nearly vertical line beyond that point.

At higher temperatures t_2 , t_3 , and so forth, the same process takes place, but with the horizontal portion of the curve becoming shorter, and liquefaction beginning with higher pressures and smaller specific volumes. Finally, with the critical temperature, the straight line vanishes at the point of inflection C . In the still higher curves the reverse curvature of the critical isothermal is gradually smoothed out, until at some much higher temperature t_n the curve becomes approximately the familiar rectangular hyperbola of the perfect gas isothermal represented by $pv = b$.

242. Areas of the phase diagram. As shown in Fig. 24, there are three different regions for temperatures lower than the critical. The all-liquid region lies to the left of the critical isothermal and of the hill-shaped curve which encloses the constant-pressure lines. The space inside this latter curve is one where liquid and vapor exist together in equilibrium, with an increasing proportion of liquid as we move toward the pressure axis with decreasing volumes. To the right and above this curve lies the gaseous area of superheated vapor. The boundary between this region and that of the liquid is marked by the critical isothermal above C , though at pressures above the critical pressure p_c , the separation is not a sharp one. Thus, starting with gas at d , if the temperature is lowered at constant pressure, we should find the specific volume steadily diminishing until at e there would be a true liquid, but no abrupt change would have occurred in crossing the critical isothermal, as would have been the case at a pressure lower than p_c .

243. Experimental demonstration of the critical point. It is possible to pass a liquid and its vapor through the critical point provided the critical temperature is not too high. If a small quantity

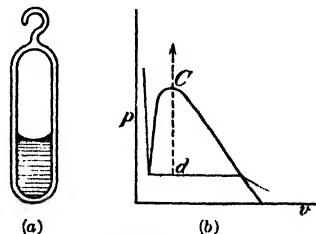


Fig. 25.

of ether ($t_c = 193.8^\circ$) is sealed up in a heavy-walled glass tube, as shown in Fig. 25 (a), so that the space above it contains only ether vapor, it may be cautiously heated to the critical temperature. The initial condition is shown at d (which

should be directly under C , as in Fig. 25 (b)), where the liquid and saturated vapor are in equilibrium. Since the total volume remains constant, a rise in temperature results in only slightly changing the relative volumes of the two states, the liquid occupying a little more and the vapor a little less, as the point C is approached. When the critical temperature is reached, the separating meniscus grows less marked and finally disappears, when there is no longer any distinction between liquid and vapor. In fact, just above C the substance partakes somewhat of the properties of both.

On cooling, the vapor-liquid becomes cloudy as it approaches C ; then the meniscus reappears, the liquid state becomes evident again, and gradually regains its original volume.

244. Van der Waals' equation. Various attempts have been made to express the isothermals we have been discussing by an "equation of state" as it is called. The equation $pv = RT/w$ ((1), Article 204), may be still further simplified by setting v_1 equal to the volume occupied by a gram molecule instead of a gram. Then $v_1 = wv$, and

$$pv_1 = RT. \quad (1)$$

This is the simplest equation expressing the condition of an ideal gas, but it is approximately true only for real gases when their temperature is far above the critical value. Near and below this temperature it is wholly inadequate.

In order to express the p , v , and T relations of a gas so as to plot isothermals which cover the process of liquefaction, certain new assumptions regarding the gas must be made. These must make allowance for the mutual attraction of its molecules, and for their finite volume which restricts complete liberty of motion. Several equations taking these facts into account have been proposed, some of them giving isothermals which agree quite closely with those obtained from experiment. But none of these is exact, and it is probable that a rigorous equation based on theoretical considerations will never be found on account of the complexity of the problem.

The most celebrated equation of state was devised by Van der Waals in 1872. It is not the most accurate, but has the advantage of simplicity, and is more easily explained than the others. It is stated thus:

$$\left(p + \frac{a}{v_1^2}\right)(v_1 - b) = RT,$$

where R has the same value as in the simpler form applicable to ideal gases, and a and b are constants depending upon the nature of the gas.

The constant b is known as the covolume, and is regarded as the smallest space the molecules of a gram molecule of gas would occupy if crowded together. The quantity a/v_1^2 must represent a pressure, for it is added to a pressure, and the sum of quantities of different dimensions is meaningless. This pressure, or force per unit area, varies inversely as v_1^2 , which would necessarily follow if it were to compensate for attractions between the molecules. When plotted, this equation gives a set of curves for different values of T , like those shown in Fig. 24, except inside the liquid-vapor region, where a curve like the dotted line replaces the straight lines ab .

245. Refrigeration. The fact that certain gases exist, under normal conditions, below their critical temperatures, and so are readily liquefied by pressure, is made use of in commercial and domestic refrigeration. Ammonia (NH_3) is one of these gases and the ammonia refrigerator is typical of several of the same kind. In Fig. 26 is shown the general scheme of such a machine. The gas is first compressed in D and forced through the valve a into the coil in B surrounded by running water, where it becomes liquid as the heat caused by the compression is removed. The valve A is then opened and the liquid rushes into the coil in C where, under greatly reduced pressure, it becomes a gas once more, and the heat of vaporization is withdrawn from the surrounding liquid. This liquid must be one which freezes only at a low temperature, and is usually a strong solution of common salt, or brine. It is made to circulate through another tank or coil, not shown, where it absorbs heat from its surroundings, cooling them below the freezing point. This results in warming the brine, which then re-enters the tank C at its upper end to be cooled once more by the vaporization of the ammonia. Finally the gas from the coil in C is drawn into the pump through the valve b and is once more compressed and ejected through a . Thus the process is a closed cycle, which may be continued with the same ammonia and brine indefinitely. Large commercial machines of this type are capable of

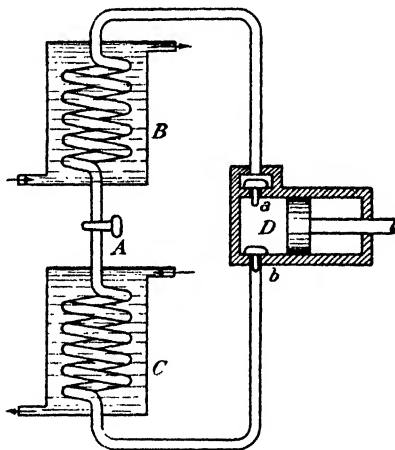


Fig. 26.

producing 30 kilograms of ice per horsepower hour expended on the pump.

246. Liquefaction of gases. The so-called “permanent gases” like oxygen and nitrogen (the chief constituents of air), and hydrogen and helium, are really no more permanent than the ammonia just referred to, but their critical temperatures are so low that pressure alone will not cause them to liquefy. The critical temperature of nitrogen is -147°C , while hydrogen must be cooled below -234° before it can be compressed into liquid form.

The method ordinarily used to liquefy gases is called a regenerative process, and was invented by Linde in 1895. The gas is first compressed in a pump to a pressure of several hundred atmospheres, and the heat of compression is removed, as in the ammonia refrigerator, by passing it through a coil surrounded by melting ice or a freezing mixture. It then passes through a spiral tube to a needle valve through which it expands to a pressure of about 15 atmospheres, and its temperature falls in the process in accordance with the principle stated in Article 201. This low-temperature gas is now passed through a second tube surrounding the first one, so as to cool its contents, as a result of which, the next installment from the pump, after passing through the inner spiral, is much colder when it expands than the first one was, and a still lower temperature is reached as a result of the expansion. Consequently the process might more properly be called a “degenerative” one, because it helps to produce progressively *lower* temperatures by a continuously reacting process of heat absorption.

Finally the gas in the inner tube reaches so low a temperature that its expansion results in liquefying a portion, and this is collected in a Dewar flask, resembling a thermos bottle. The portion not liquefied is carried back through a third tube surrounding the other two, where it expands to atmospheric pressure, thus still further cooling the gas in the two inner tubes.

SUPPLEMENTARY READING

Saha and Srivastava, *A Textbook of Heat* (Chap. 6), Indian Press, Allahabad, 1931.

CHAPTER 20

Heat and Energy

247. Heat as a form of motion. In studying the kinetic theory of gases, we saw that heat consists in the total kinetic energy of their molecules. This view of heat however is a comparatively recent one. During the first part of the 18th century it was supposed that when matter burned it liberated an elementary substance called phlogiston (fire substance) which the matter had contained in a latent form. This theory was later discarded, and a different and more satisfactory one was developed which accounted for heat, whether associated with combustion or not, by supposing it to be an imponderable fluid called "caloric," a name first used by Lavoisier in 1789. When this fluid entered a body, the body became warm, and when it flowed out the temperature of the body fell. The heat of friction was explained by assuming that when two bodies were rubbed together, small particles were rubbed off, and that this product of abrasion had less capacity for caloric than the original substance. Therefore some of its caloric fluid was liberated from the latent form and became sensible heat.

The first to doubt this theory was Count Rumford. He was born in Woburn, Massachusetts, in 1753 and named Benjamin Thompson. After a most eventful career he entered the service of the Elector of Bavaria, where his advancement was rapid. He ultimately became Minister of War and was made a count, taking his title of Rumford from the name of his wife's home, Rumford, now Concord, New Hampshire.

As Minister of War, Rumford was associated with the arsenal at Munich, and was impressed with the great amount of heat developed in boring cannon. This seemed to him too great to be accounted for by the caloric hypothesis. Moreover he found by experiment that the metallic chips produced by the boring had the same capacity for heat as the original metal. Several experiments, in which the heat developed was measured, only confirmed his doubts, and he finally concluded that it could not be a material substance, but that motion alone was a possible explanation of heat. This however did not convince the supporters of the caloric theory, because Rumford

had not actually proved that heat could be produced in unlimited quantities by friction, although he believed this to be the case.

The matter was finally settled by Sir Humphry Davy, who in 1799 performed a conclusive experiment which put an end to the caloric hypothesis. This consisted in rubbing two pieces of ice together, when the heat of friction melted them and produced water as the product of abrasion. The calorists admitted that water contained more heat than ice at the same temperature, because when water was formed from melting ice, it absorbed a great deal of heat without rising in temperature. Therefore in this case the product of rubbing was a substance that contained more heat than the body from which it was derived, and there must have been an increase in the total amount of heat as a result of the friction.

248. Heat a form of energy. Although Davy's experiment was a conclusive proof that heat was not a fluid contained in matter, the alternative view proposed by Rumford, that it was a mode of motion, was not wholly satisfactory, and it was not until 1842 that Mayer, a German physicist, definitely stated the equivalence of heat and energy. Somewhat earlier than that, in 1840, Dr. James P. Joule had begun experiments in Manchester on the relationship between the energy of chemical reactions and electrical energy, and was led to a similar conclusion. In 1843 he experimented with a dynamo whose current heated a conductor, and actually measured the work done in driving the dynamo, which he compared with the heat evolved. This comparison gave 838 foot-pounds of work as the mechanical equivalent of a British thermal unit, and Joule announced before the British Association his conclusions regarding the "convertibility of heat and mechanical power" in either direction.

249. The mechanical equivalent of heat. In spite of a very skeptical reception of his theory, Joule continued his experiments in the same direction with a view to obtaining even more conclusive evidence, as well as more accurate quantitative measurements. In 1845 he announced the result of an experiment in which measured mechanical work was directly converted into measured heat. From this he calculated that the ratio of these values, or the **mechanical equivalent** of heat, was 798 foot-pounds per B.t.u. More recent measurements make this important constant 777.9, which corresponds to 4.185 joules per gram calorie, and is the accepted value today, though Birge in 1929 announced one more significant figure giving the value of 4.1852. This factor for converting heat units into units of work is usually denoted by the letter J , so that we may write $W = JH$, where

H is the heat developed by the work W . The numerical value of J differs of course, according to the units used to measure heat and mechanical energy, and the following table gives the conversion factor

	Gram calories	B.t.u.	Joules	Foot-pounds
Gram calories.....	1	3.968×10^{-3}	4.185	3.087
B.t.u.	252	1	1055	777.9
Joules.	0.2389	9.482×10^{-4}	1	0.7376
Foot-pounds.....	0.3240	1.286×10^{-3}	1.356	1

for several cases, as well as its reciprocal, to be used in converting work into heat. Thus, to convert B.t.u. to joules, use $J = 1055$, while to convert joules to B.t.u., multiply the mechanical work by 9.482×10^{-4} .

250. Joule's experiment. The classic experiment by which Joule determined the value of J consisted in driving a paddle wheel by means of descending weights which hung from cords wrapped around a spindle on the vertical axis of the paddles. When these were made to rotate, they churned up the water in a calorimeter, and the heat developed by the friction of the water was measured in British thermal units, as well as the foot-pounds of energy developed by the descending weights. A comparison of these quantities gave the desired equivalent.

In a later and better form of the apparatus, ten paddles, in two sets of five each, rotated about a vertical axis within a calorimeter equipped with four stationary vanes to increase the churning effect on the water, which resulted in a large torque on the containing vessel. The latter was mounted so that it could rotate almost without friction, but rotation was prevented by two cords supporting weights, as shown in Fig. 27 (a). The shaft E which carried the paddles was driven by the operator, who turned the grooved wheel A belted to a smaller wheel B . The paddles d (Fig. 27 (b)) were enclosed in the calorimeter C , which was supported by a float contained in the tank S partly filled with water, and when rotated they set the water in C in motion, thus acting upon the vanes a to produce a torque upon the calorimeter case. This torque increased with the speed of rotation, and the operator turned A with the speed just necessary to keep the weights in equilibrium off the ground. When the speed

of rotation was exactly right, the torque on C was just equal and opposite to the couple due to the weights m . This couple is computed from the radius r of the circular groove around which the cords were wrapped, and the force mg exerted by each weight, giving

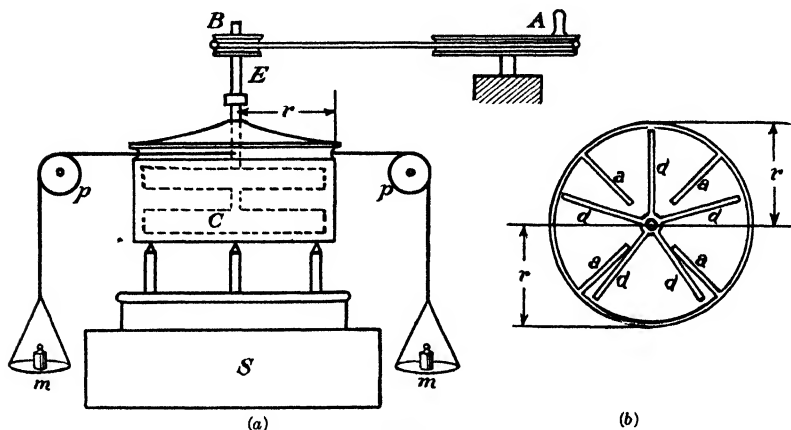


Fig. 27.

$L = 2mgr$. But this is equal to the torque exerted on C by the paddles, and the work they perform when turned through an angle θ is $L\theta$. After n revolutions, $\theta = 2\pi n$; therefore the total energy is $W = 4\pi nmgr$.

The heat developed was measured by observing the rise of temperature of the water, paddles, and calorimeter, and calculated in the usual manner from the known specific heats of all the materials affected, so that the heat gained is given by $H = (t_2 - t_1)\Sigma sm$. Then the equation of equivalence, $W = JH$, becomes

$$4\pi nmgr = J(t_2 - t_1)\Sigma sm,$$

from which J may be calculated.

A number of other mechanical methods for finding Joule's equivalent have been used, and it has also been found from the heat developed by an electric current flowing through a wire, and from the heat due to eddy currents in a mass of metal, caused by a rotating magnetic field. A purely calorimetric method was devised by Mayer,* who used equation (2) of Article 214, $J(s_p - s_v) = R/w$. Thus, if we take the specific heats of hydrogen, $s_p - s_v = 3.39 - 2.40 = 0.99$. Its molecular weight is 2.016, and $R = 8.313 \times 10^7$. Therefore $J = 8.313 \times 10^7 / (2.016 \times 0.99) = 4.16 \times 10^7$ ergs per calorie.

251. The first law of thermodynamics. The science which deals with the relations between heat and other forms of energy is known as **thermodynamics**. The principle of the equivalence of heat and energy, though it is incapable of a theoretical proof, is now accepted as a universal law, and is called the "first law of thermodynamics." As this law refers only to transformations of *heat* into other forms of energy, and vice versa, it may be stated as follows: *When heat is transformed into some other form of energy, or some other form of energy is transformed into heat, there is an exact equivalence between the energy which disappears and that which is produced at its expense.*

252. Energy transformations. The transformation of mechanical energy into heat is a very easy and natural process, and is normally one hundred per cent efficient. It occurs whenever friction develops, as when two rough objects are rubbed together, a process which converts all the applied energy into heat. The heat due to the compression of a gas is another illustration of complete conversion. Such transformations are so natural, indeed, that it is difficult to avoid them. We grease machinery to reduce the heat loss of friction, and in transmitting electricity use wires of large section to minimize their resistance to the electric current, which causes a loss of energy through its transformation into heat.

But the conversion of heat into higher forms of energy is essentially artificial, and is never complete, for even in ideal steam engines, the efficiency falls far short of unity. To effect such a transformation, some kind of engine is necessary in which thermal energy, or its equivalent in the form of chemical combination, results in performing some kind of useful work. Even water motors are types of such a conversion, for waterfalls are due to the sun's heat, which causes water to evaporate from the sea to be condensed later as rain or snow and ultimately to flow as rivers back into the sea.

We may sum up the foregoing in the general statement that all forms of energy tend toward heat, the lowest form of energy, and that such transformations may easily be made complete, while the reverse process is unnatural, incomplete, and requires special apparatus. But, complete or incomplete, the amount of energy which is converted into another form reappears in that form in an exactly equivalent amount.

253. Intrinsic energy. When heat is converted into work, the work is done at the expense of a portion of the heat energy which disappears in the process, in accordance with the first law. The remainder, which, as we have seen, cannot be so transformed, results

either in a rise of temperature of the working medium, or in a change of its molecular structure, or both. This molecular change might mean the vaporization of water, or the melting of ice, when heat apparently disappears, or becomes "latent." In either case the body has gained **intrinsic energy**, which it possesses by virtue of its temperature or molecular structure. That is, its internal kinetic or potential energy has been increased.

If we represent intrinsic energy by the letter U , we may express the transformation of heat into mechanical energy or vice versa by the equation

$$\Delta H = \Delta W + \Delta U,$$

where the symbol Δ means a change in each of the three items. This is a general statement of the first law of thermodynamics, and is always applicable. When the working medium receives heat, ΔH is positive. When it does work, ΔW is positive, and a positive value of ΔU means a gain in intrinsic energy. If ΔU is small compared to ΔW , the conversion is highly efficient. If it is large, the efficiency is correspondingly reduced. When no external work is done in the process, as when a body is heated in a vacuum, $\Delta H = \Delta U$. It is also possible for a body to do work at the expense of its intrinsic energy without the application of external heat, as in an explosion resulting from chemical recombination. Then $\Delta H = 0$ and $\Delta W = -\Delta U$. Finally, when work produces heat, $\Delta U = -\Delta W$, which means that mechanical energy has been wholly converted into a change of intrinsic energy, as when ice is melted or water heated by friction.

254. Isothermal and adiabatic processes. When an ideal gas is compressed or allowed to expand **isothermally**, that is, without change of temperature, it follows Boyle's law, and the pressure-volume curve is the familiar rectangular hyperbola which approaches the axes as asymptotes according to the equation $pv = b$. This is essentially an unnatural process, for as we have seen, compression tends to raise the temperature of a gas while expansion tends to lower it. Therefore, to maintain a constant temperature, the heat of compression must be removed, as by a water jacket, or during expansion the tendency to cool must be counteracted by supplying heat from outside.

But instead of keeping the temperature constant, we may compress a gas so that none of the heat of compression is lost, or allow it to expand without influx of external heat. Such a process is called **adiabatic**, meaning *not flowing through*, and is more natural and more easily realized than an isothermal process. If the cylinder in which the compression or expansion is to take place is surrounded by a

perfectly insulating medium, no heat can flow out from or into the working medium, and the process is adiabatic. The equation of such a process may be shown to be $pv^\gamma = b$ for ideal gases, where γ is the ratio s_p/s_v of the specific heats of the gas, and is always greater than unity.

Actual expansions or compressions are ordinarily neither exactly adiabatic nor isothermal, for it is difficult to maintain the temperature constant, and though insulation is a much easier condition to achieve, it can never be made absolutely perfect. But a very sudden change in volume is necessarily almost an adiabatic process, because there is not enough time for a perceptible flow of heat either in or out. On the other hand, a very slow change can be made almost isothermal, as the medium under these conditions maintains the temperature of its surroundings without especial precautions.

255. Adiabatic curves. The curve of an adiabatic expansion or compression is not hyperbolic, but as can be seen from its equation $pv^\gamma = b$, small changes in the volume involve correspondingly larger changes in the pressure. This is shown in Fig. 28, where the solid adiabatic curve CA is steeper than the dotted isothermals.

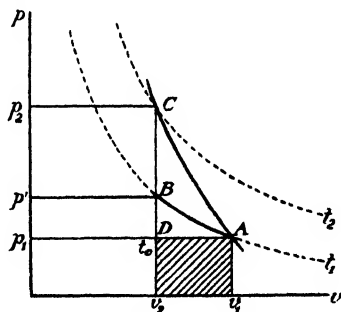


Fig. 28.

The greater steepness of adiabatics is easily understood from purely physical considerations. Thus suppose a gas at the point A , whose specific volume and pressure are v_1 and p_1 . Let it be compressed isothermally at the temperature t_1 to the volume v_2 at B by removing the heat of compression during the process. This causes the pressure to rise to p' . Then starting at the same point A , let the gas be compressed adiabatically to the same volume as before at C . During this process the gas is insulated, and the temperature rises to a higher value t_2 , so the adiabatic cuts the v_2 line at a higher pressure level, p_2 , on the t_2 isothermal. Therefore the curve is steeper.

256. Areas on the p - v diagram. Whenever points on a plane are referred to a system of rectangular co-ordinates, areas on such a plane are proportional to the products of these co-ordinates.

On the pressure-volume diagram, areas measure the product pv , which has been shown to be work. So in Fig. 29, if a gas expands from the point (p_1, v_1) to (p_2, v_2) , its smooth curve of expansion may be

broken up into minute steps, each indicating a change of volume Δv and an amount of work equal to $p\Delta v$. The total work done is the sum of all such steps, and if they are made vanishingly small, this is identical with the area between the curve, the axis of volumes, and the v_2 and v_1 lines.

One illustration of the use of areas on the p - v diagram is in connection with the specific heats of a perfect gas. In Fig. 28, a gas

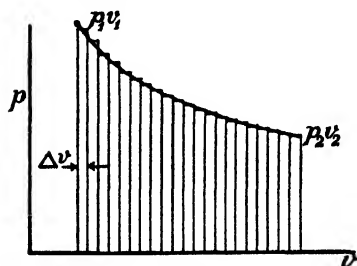


Fig. 29.

heated under constant pressure from the temperature t_0 at D to t_1 at A , expands from v_2 to v_1 in the process and does work. The amount done is proportional to the shaded area directly under the p_1 line. Let t_1 be one degree higher than the temperature t_0 . Then the work represented by the shaded area accounts for the difference between

the specific heat at constant pressure and that at constant volume, for when heated at constant volume to t_1 , the gas moves up the v_2 line from D to B . This involves no work done by the gas and the specific heat increases only its intrinsic energy.

257. Isothermal and adiabatic elasticity. The elastic modulus of an ideal gas at constant temperature is nearly equal to the pressure, provided the change in volume is small. This is proved as follows: Suppose a small increase of pressure Δp results in a correspondingly small decrease in volume Δv ; then the product of the new pressure, $p + \Delta p$, times the new volume, $v - \Delta v$, equals the original product pv ; then

$$(p + \Delta p)(v - \Delta v) = pv,$$

$$\text{and} \quad pv + v\Delta p - p\Delta v - \Delta p\Delta v = pv.$$

By canceling and collecting terms, we obtain

$$v\Delta p = \Delta v(p + \Delta p),$$

$$\text{or} \quad v \frac{\Delta p}{\Delta v} = p + \Delta p.$$

But the elastic modulus B is defined as the ratio of the change in pressure to the relative change in volume, or

$$B = \frac{\Delta p}{\Delta v/v} = v \frac{\Delta p}{\Delta v},$$

which is therefore equal to $p + \Delta p$, as proved above. The small change in pressure, Δp , may be neglected in comparison with p ; therefore $B_t = p$ approximately, where the subscript t has been introduced to indicate the fact that the temperature was kept constant, and that B_t is the *isothermal modulus of elasticity*.

If the compression is adiabatic,

$$(p + \Delta p)(v - \Delta v)^\gamma = pv^\gamma.$$

Expanding the second parentheses by the binomial theorem, we have

$$(p + \Delta p)(v^\gamma - \gamma v^{\gamma-1} \Delta v + \dots) = pv^\gamma,$$

where the terms involving higher powers of Δv are dropped as being negligibly small. Then

$$pv^\gamma - \gamma pv^{\gamma-1} \Delta v + v^\gamma \Delta p - \gamma v^{\gamma-1} \Delta v \Delta p = pv^\gamma.$$

Dividing by $v^{\gamma-1}$ and transposing, we obtain

$$v \Delta p = \gamma(p \Delta v + \Delta v \Delta p),$$

or
$$v \frac{\Delta p}{\Delta v} = \gamma(p + \Delta p),$$

whence $B_\phi = \gamma p$, when Δp is negligible compared to p , and the subscript ϕ indicates an adiabatic modulus, which is greater than B_t since $\gamma > 1$.

258. Free expansion of gases. We have seen that when a gas expands against some opposition and does work, it will be cooled if no heat flows in to warm it. But what should we expect if it expands in an insulated enclosure without doing work? To answer this question let us use the first law of thermodynamics expressed by $\Delta H = \Delta W + \Delta U$. Here both ΔH and ΔW are zero, because no heat flows in or out, and no work is done, by hypothesis. Therefore $\Delta U = 0$, which means no change in intrinsic energy and *no change in temperature*. To see if it were

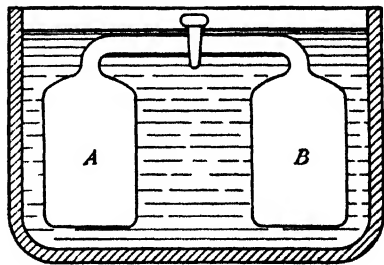


Fig. 30.

true that a freely expanding gas was not cooled, Joule in 1845 performed the celebrated experiment illustrated in Fig. 30. A gas compressed in chamber A was allowed to expand into a partial vacuum in B. The whole apparatus was immersed in water con-

tained in a double-walled tank which acted as a fair insulator. In such a process no *external* work is done by the expanding gas, and since the process is almost wholly adiabatic, $\Delta H = 0$ approximately. The experiment with a real gas seemed to be in accordance with this theory, for Joule was unable to detect any change of temperature in the surrounding water.

When the vessels *A* and *B* were immersed in separate vessels, Joule found that the one containing *A* showed a fall of temperature due to the expansion, while that which contained *B* rose in temperature by an apparently equal amount because of the compression taking place there. But as the two effects seemed to balance each other, Joule again concluded that the intrinsic energy of the gas as a whole had not changed and was therefore independent of the volume it occupied, so that free expansion not involving *external* work was an isothermal process. This is now known to be true only of ideal gases, but it is only a very rough approximation in the case of a real gas, a fact which escaped Joule because the temperature changes in the gas were not enough to affect appreciably the rather large body of water around it.

259. The porous-plug experiment. In 1852 William Thomson (later Lord Kelvin) collaborated with Joule in a much more delicate test of the foregoing principle. They forced a gas under pressure slowly through a plug of cotton wool and measured its temperature before and after the expansion, which was practically adiabatic. Just outside the plug there was always a certain amount of cooling due to the eddies in the escaping gas, because their production entailed a loss of intrinsic energy. But at a sufficient distance from the plug, where the eddies had subsided, Joule and Thomson found a definite lowering of the temperature of all gases but hydrogen, which exhibited an unexpected rise.

In the case of an ideal gas, such an experiment, ignoring the temporary cooling near the plug, should result in no change of temperature of the expanding gas, because the work of compression on the high-pressure side of the plug would equal the work done by expansion on the low-pressure side; therefore $\Delta W = 0$, and as the process is adiabatic, ΔH is zero also, and there could therefore be no change in U . Since such a change was observed, the only possible conclusion was that ΔW could not be zero. In other words, the gas did not behave like an ideal one, and must have done *internal* work during the expansion. Then ΔW is positive and ΔU negative, showing that the internal mechanical work is done by the gas at the expense of intrinsic energy, so the temperature must fall. But if, as in the case of

hydrogen, expansion should involve internal work done *on the gas*, then ΔW is negative, ΔU positive, and the temperature rises.

Further study of this anomalous behavior of hydrogen showed that if its temperature were below -80°C , it behaved like other gases, while they in turn exhibited the unexpected heating effect if their temperatures were above a certain critical value known as the **temperature of inversion**.

The Joule-Thomson cooling plays an important part in the liquefaction of gases, for it becomes increasingly important under high pressure and at low temperature. As this condition is approached, the intrinsic energy of the gas is more and more influenced by its volume, because of the increasing attraction between the molecules as they are crowded closer together. If then the gas is allowed to expand adiabatically, the work needed to separate the molecules against their mutual attractions is taken from their kinetic energy, which results in a lowering of the temperature.

260. Carnot's cycle. The problem of how to convert heat into mechanical energy as efficiently as possible led Sadi Carnot, a French engineer, in 1824 at the age of 28, to devise an imaginary heat engine that has since become famous, because it stands for the ideal which may be approached in practice but never surpassed. In this hypothetical machine, a fluid, or "working substance," which expands when heated, is put through a cycle of operations, at the end of which it is in its original state. This is known as a **closed cycle**, and may be represented by the series of changes which the water of a condensing engine undergoes; namely, vaporization, expansion, exhaust, condensation, and return to the boiler through the feed pump. But ideally, Carnot's cycle, in addition to being closed, is also **reversible**. That is, it can be traced through in either direction, which is impossible in most cycles. In the one just mentioned, for instance, reversibility would mean that water from the boiler must go back into the feed pump at a temperature below the boiling point, while the heat it has lost remains in the boiler at boiler temperature, which is impossible.

Carnot supposed a cylinder fitted with a frictionless piston which encloses a gas or vapor. The walls of the cylinder and the piston are insulators, but the base of the cylinder is a perfect conductor of heat. Thus the gas may be heated by an inward flow through the base, or cooled by an outward flow, or the base may be protected by an insulator so that no flow of heat takes place. This arrangement is indicated in Fig. 31, where *I* is an insulating stand upon which *C*

may be placed when no flow of heat is desired during an adiabatic process. A generator, G , supplies heat in unlimited quantity at some fixed temperature without becoming cooler as it delivers it; and R is

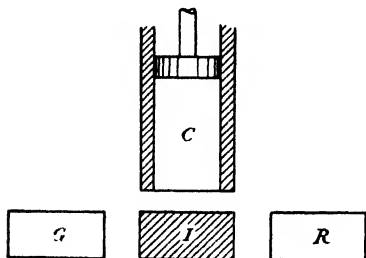


Fig. 31.

a refrigerator supposed able to absorb heat at a constant temperature without becoming warmer. When C is resting on I , an expansion or a compression is adiabatic, while when resting on G or R , both processes are isothermal.

Now, without further reference to this imaginary mechanism, suppose the gas starts at a (Fig. 32) with a pressure p_1 , specific volume v_1 , and temperature t_1 . Let it expand isothermally to b , receiving heat as it does so from the generator. During this process, it does $+W_1$ units of work, measured by the area $abb'a'$. Next the cylinder is insulated, and the gas allowed to expand adiabatically from b to c , doing work $+W_2$, proportional to the area $bcc'b'$. In this process its temperature falls to t_2 , so that if it is now compressed isothermally, it follows the t_2 isothermal and delivers heat to the refrigerator in so doing. This compression requires work, and an amount $-W_3$, proportional to $cc'd'd'$, is done upon it. The point d is not an arbitrary one, for it must be chosen so that the original condition may be reached by adiabatic compression. The final step then consists in compressing the gas adiabatically, when an amount of work $-W_4$, proportional to the area $dd'a'a'$, is done upon it.

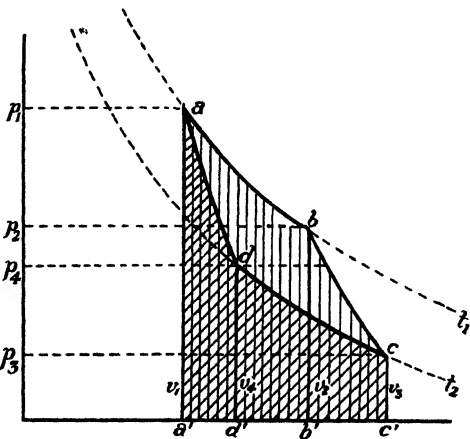


Fig. 32.

The net result of the cycle is $W_1 + W_2 - W_3 - W_4$, and this is proportional to the area $abcd$, which has only vertical hatching in the diagram. The crosshatching indicates areas where work

done by the gas is neutralized by work done upon it. The area $abcd$ represents work done by the gas, because the areas representing positive work are clearly larger than those representing negative work.

The Carnot cycle is perfectly reversible, for if gone through in the reverse direction, the adiabatic expansion ad delivers work proportional to $add'a'$, dc delivers $dcc'd'$, cb requires $ccb'c'$, and ba requires $baa'b'$. The net result is now $W_3 + W_4 - W_1 - W_2$, where the subscripts refer to the same areas as above. Therefore, since the areas representing negative work (done on the gas) are now the largest, the net result is negative, and the amount of energy required to make the engine perform this reversed cycle is proportional to $abcd$.

261. Thermal changes in Carnot's cycle. After the direct cycle is completed, it is clear that the generator G has lost the heat delivered to the medium during the isothermal expansion, while R has gained heat during the isothermal compression. Carnot supposed that heat was like a fluid and indestructible, so that in such an engine it did work as water does in running a mill wheel, by descending from the high temperature level t_1 of the generator to the lower level t_2 of the refrigerator, where he supposed *all* the calories taken from G were dumped, as into the tailrace of a water mill. But this hypothesis is contradictory to the first law of thermodynamics, because when work is done, an equivalent amount of heat must disappear. Therefore, if we represent the heat taken from G by H_1 , and that delivered to R by H_2 , the total work is given by $W = J(H_1 - H_2)$.

If the engine is run backward, the heat is removed from R during the expansion dc , and delivered to G during the compression ab . But since work is done *on* the gas instead of *by* it, heat must be created, so that more heat is delivered to G than was removed from R , and we have $-W = J(-H_1 + H_2)$, where $H_1 > H_2$.

262. Refrigeration and heating by Carnot's cycle. The reversed Carnot cycle is really a refrigerating machine, for R is constantly losing heat as a result of the work done upon the gas. Indeed, a similar device is actually used in one form of commercial refrigerator. But on the other hand, G gains more heat than R loses, so that the Carnot engine run backwards might be regarded as a device for lifting heat from a low temperature to a higher one, while increasing the total amount on the way, at the expense of mechanical energy supplied to it. Theoretically this would be an ideal way to heat a house, by taking heat from the cold outer air, and dumping an increased amount, at the desired room temperature, into the house.

263. The second law of thermodynamics. This important law is a statistical principle based on overwhelming probability. It can be formulated in a variety of ways, but all of them really mean that *heat tends to flow from higher to lower temperatures.*

Perhaps the most useful statement of the second law is due to Clausius, who announced that: "*It is impossible for a self-acting machine, unaided by any external agency, to convey heat from one body to another at a higher temperature.*" That this is only statistically true

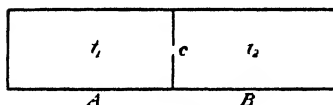


Fig. 33.

may be shown as follows: Let us imagine two chambers A and B separated by a diaphragm with a small window c , as in Fig. 33. Let each chamber contain gas at the same pressure but at different tem-

peratures, such that t_1 is higher than t_2 . If the pressure is very low, indicating a relatively small number of molecules in each compartment, it is easy to imagine the chance that occasionally a particularly slow-moving molecule in A might pass through the window into B , while a particularly fast one in B might thus get into A . If the more probable contrary exchange did not occur during the time we are considering, the result would be to make A warmer than before, and B cooler. But when countless millions of molecules are involved, this chance is vanishingly small, and though an occasional accident like the above is bound to happen, it is overwhelmed by the vast majority of cases where B rises and A falls in temperature.

Maxwell suggested that intelligence might defeat the second law, and supposed an intelligent "demon," sitting at the window c , controlling a frictionless trap door over the opening. With this he could "sort the molecules," letting only exceptionally fast ones through from B to A , and exceptionally slow ones through from A to B , thus causing heat to flow "up hill" without expenditure of energy. It has even been suggested that certain inexplicable phenomena connected with living organisms may be accounted for on the assumption that living cells behave like Maxwell's demon.

The second law is vitally concerned with the theory of thermodynamics, and has been taken for granted in much of what we have already learned about heat and energy in this chapter, as when it was assumed that heat flows unaided from hot to cold, and not from cold to hot. But there are many more uses for the second law in the more advanced development of thermodynamics, and a number of ways of stating it, besides those already mentioned.

264. Efficiency of reversible cycles. In the Carnot cycle, the heat H_1 supplied at the higher temperature t_1 is the thermal energy supplied to the engine, while the output is the energy equivalent of that which disappears as heat and is transformed into mechanical work, or $H_1 - H_2$. Therefore the efficiency, output divided by input, is given by

$$e = \frac{H_1 - H_2}{H_1}.$$

Now it was proved in Article 200 that the mean kinetic energy of the molecules of a gas is proportional to the absolute temperature, so that $W \propto T$. But the first law of thermodynamics states that the thermal energy H is proportional to W ; therefore $H \propto T$. Consequently $H_2/H_1 = T_2/T_1$; whence by "division"

$$e = \frac{H_1 - H_2}{H_1} = \frac{T_1 - T_2}{T_1}.$$

The fact that the efficiency of a Carnot engine is given by this simple relation between the two extreme temperatures of the cycle is of great importance, and is true of all other reversible heat cycles. It may also be proved that no other cycle operating between the same temperature limits can yield as high an efficiency as that of a reversible cycle. This is known as Carnot's principle. Carnot efficiency is an unattainable ideal which, for any conversion of heat into mechanical energy, can be only approximated in practice, since real cycles are necessarily more or less irreversible.

It is evident that the Carnot efficiency increases as we raise the temperature T_1 of the generator, or lower T_2 of the refrigerator, but it can never reach unity unless the refrigerator could be maintained at the absolute zero. Then $T_2 = 0$, and $e = T_1/T_1 = 1$.

265. Irreversible cycles. Whenever heat flows unaided from a hotter to a colder body, or motion develops heat of friction, the process is irreversible. As these processes occur in all real heat engines, the engines necessarily operate in irreversible cycles. If, for instance, the cycle involves the condensation of steam in contact with cold water pipes, the reverse process would call for a return of the heat of vaporization from those pipes, a flowing from cold to hot, which is impossible. The heat caused by friction also represents an irreversible transformation. Thus the rotation of a wheel on its axle results in heating the bearings, but, by turning the wheel the other way, we cannot restore the lost mechanical energy that this heat represents.

The irreversible loss of heat up the chimney of a power plant, or the heat which escapes into the boiler room, helps to prevent the realization of ideal efficiency. But care should be taken to avoid the common error that if there were no such losses and no friction, one hundred per cent efficiency could be obtained. This of course is not the case, for the Carnot engine, because of its perfect reversibility, and because no losses of any sort are contemplated, puts an upper limit to the performance of any heat engine. Therefore an engine operating between, say, 200°C in the boiler, and 20°C in the condenser can never exceed an efficiency of

$$\frac{T_1 - T_2}{T_1} = \frac{473^{\circ} - 293^{\circ}}{473^{\circ}} = 38 \text{ per cent, approximately.}$$

266. Indicator cards. The variation of pressure and volume within the cylinder of a steam or internal combustion engine may be represented in diagrams called indicator cards. These are not true cycles

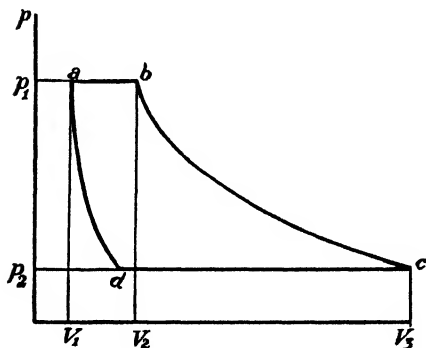


Fig. 34.

in the sense in which we have been using that word, but they are extremely useful in studying the performance of such engines, and in determining the power they develop. A typical steam-engine indicator card is shown in Fig. 34. The steam is admitted to the cylinder, where the "clearance" volume is V_1 , under boiler pressure p_1 . As the piston goes out, the increasing space is

filled with live steam from the boiler at the same pressure, until at b , the admission valve is closed. After this "cutoff" it expands more or less adiabatically to c at the end of the stroke with a total volume V_3 . The exhaust valve then opens and the piston starts on its return stroke against a pressure p_2 which is that of the condenser in condensing engines, but is otherwise atmospheric, or nearly so. At d the valve is closed, and the piston compresses the remaining steam along the curve da , so that at the end of the stroke there is a cushion of steam filling the clearance, and already raised to boiler pressure.

As in cycle diagrams, the work done is proportional to the closed area, but the efficiency of the engine as a whole, including boiler,

condenser, and so forth, cannot be obtained from it. The horsepower developed is calculated from the work per stroke, in foot-pounds, multiplied by the number of strokes per second and divided by 550. This is known as the "indicated horsepower," and is always somewhat greater than the actual power output because of friction and other losses.

SUPPLEMENTARY READING

H. A. Perkins, *An Introduction to General Thermodynamics* (Chap. 2), Wiley, 1916.

Enrico Fermi, *Thermodynamics* (Chapters 2, 3), Prentice-Hall, 1937.

PROBLEMS

1. Calculate the amount of heat developed when a mass weighing 80 kg is moved 20 m along a rough horizontal surface where the coefficient of friction is 0.7. *Ans.* 2622 calories.

2. A bullet weighing 6 oz. and moving with a speed of 1800 ft./sec. strikes a target which stops it completely. How much heat is developed? *Ans.* 24.26 B.t.u.

3. Ten grams of air are heated from 0°C to 60°C under atmospheric pressure. Calculate the change in volume, the work done, and the heat which is thus converted into work, taking the density of air as 1.3 g per liter at 0°C . *Ans.* 1690 cm^3 ; 171.25 joules; 40.9 calories.

4. The specific heat of air at constant pressure is approximately 0.24. Calculate the heat supplied in Problem 3, the gain in intrinsic energy, and the specific heat at constant volume. *Ans.* 144 calories; 103.1 calories; 0.17 calories per gram.

5. Calculate the efficiency of energy conversion in Problems 3 and 4. *Ans.* 28.4 per cent.

6. The density of saturated steam at 100°C is 0.606 g per liter. How much heat does the expansion from water represent, and what fraction of the total heat of vaporization? *Ans.* 39.94 calories; 7.4 per cent.

7. Saturated steam at 100°C in a large and thermally insulated cylinder pushes a frictionless piston against a pressure of 74 cm of mercury so as to increase its volume by 5000 cm^3 . How many grams are condensed? *Ans.* 0.219 g.

8. Calculate the efficiency of an ideal reversible engine operating between the temperatures of 0°C and 141°C . *Ans.* 34 per cent.

9. If the heat supplied to the engine in Problem 8 is produced by the combustion of one pound of coal per hour, and if this coal contains 3.5×10^6 calories per pound, what horsepower is developed? *Ans.* 1.86 hp.

10. A steam engine demands the combustion of 1.5 lb. of coal (specified in Problem 9) per horsepower hour. What is the thermodynamic efficiency of engine and boiler? *Ans.* 12.2 per cent.

* 11. A 500 hp. steam turbine and its boiler have an efficiency of 18 per cent. How much coal (specified in Problem 9) is burned in 24 hours? *Ans.* 6.1 tons.

* 12. A reversible engine develops 8 hp. The high temperature of the cycle is 140°C and the low temperature is 20°C . How many B.t.u. are drawn from the generator per second, and how many are delivered to the refrigerator? *Ans.* 19.47 B.t.u.; 13.81 B.t.u.

* 13. A reversible engine is run backwards to operate a refrigerator maintained at -15°C in a room whose temperature is 20°C . How many kilowatt-hours are needed to form 500 kg. of ice at -15°C from water at 20°C ? (NOTE: The specific heat of ice is about half that of water.) *Ans.* 8.48 kw. hr.

CHAPTER 21

Solutions

267. Solutions in general. A true solution is a homogeneous mixture of two or more substances whose proportions may be varied within certain limits, and which do not naturally become separated. There are nine possible types of solutions, all of which are known. These are:

1. Solutions of
 - (a) solids in solids (e.g., alloys)
 - (b) liquids in solids (e.g., amalgams)
 - (c) gases in solids (e.g., hydrogen in palladium)
2. Solutions of
 - (a) solids in liquids (e.g., salt in water)
 - (b) liquids in liquids (e.g., water in alcohol)
 - (c) gases in liquids (e.g., carbonated water)
3. Solutions of
 - (a) solids in gases (e.g., iodine vapor in air)
 - (b) liquids in gases (e.g., moist air)
 - (c) gases in gases (e.g., dry air)

The last of these is the most ideal, for gases mix with each other in all proportions, and the laws of such mixtures have already been discussed.

The solution of one liquid in another is sometimes as perfect as that of gases, as when alcohol and water are mixed, for one dissolves the other to any extent, but many liquids mix only in limited amounts, and some practically not at all, like oil and water. An illustration of limited solubility is given by ether and water. Water dissolves freely up to about 10 per cent of ether, while ether can hold only about 3 per cent of water in solution. In the first case, water (the larger component) is commonly called the *solvent* and then ether is called the *solute*, but in the second case ether would naturally be called the solvent and water the solute. However, it should be understood that solubility is mutual, and it is just as logical to say that water dissolves in a lump of sugar, as to make the more usual statement. In what follows, then, the word **solvent** will in general

be used for convenience to denote the constituent of larger quantity, and **solute** the constituent of smaller quantity. But in the case of solids or gases in *liquid* solutions, the liquid is usually considered the solvent regardless of quantity.

268. Saturation. In such solutions as alcohol and water, there is no limit to the possible proportions, but when such a limit exists, a condition known as saturation is arrived at as more and more of the solute is added to the solvent, and further addition is, so to speak, useless. A **saturated solution**, then, may be defined as *one whose concentration is unchanged by further addition of the saturating substance, or solute.*

In the case of ether and water, an excess of ether when it is added to water, regarded as the solvent, results in forming merely a layer of undissolved ether, which floats on the surface without altering the concentration, while excess of water added to ether produces a layer of undissolved water at the bottom. If either of these systems is shaken up thoroughly, a condition of equilibrium results where the water layer is saturated with respect to ether, and the ether layer with respect to water.

269. Gases dissolved in liquids. This case is somewhat more complicated, for the amount dissolved depends not only on the nature of the gas and liquid, and the temperature (which affects most solutions), but also upon the pressure of the gas. Water dissolves air in limited amounts, and it is this air that enables fish to live in water. They would drown for want of air to breathe, in water that had lost its dissolved air by boiling. Carbonic acid gas (CO_2) is relatively soluble in water, forming a solution commonly, but erroneously, called soda water. In fact, all effervescent beverages are solutions of this gas under pressure. Ammonia (NH_3) dissolves with immense freedom in water, forming a chemical compound known as ammonium hydroxide, the "aqua ammonia" of commerce. Water holds a dissolved gas better at low temperatures than high, which accounts for the fact that a cold bottle of soda water may be uncorked with less risk of spurting, than a warm one; while the bubbling of a glass of such a liquid may be greatly stimulated by warming it.

270. Solids as solvents. The solution of one solid in another is best accomplished when at least one of them is in the molten state. Thus carbon is readily dissolved in molten iron, forming steel. The various well-known alloys, such as brass (a mixture of copper and zinc), or bronze (a mixture of copper and tin), are made by mixing both metals as liquids. An illustration of a liquid dissolved in a solid is that of mercury, which alloys readily with solid gold, forming a

solution. This is called an amalgam, when the amount of mercury in solution is relatively small. Gases also may be absorbed by solids, forming a solution. Palladium absorbs hydrogen so vigorously as to grow red hot while its volume increases very perceptibly. Platinum also has this power, and one form of cigar lighter depends upon the adsorption of the vapor of wood alcohol by a platinum surface.

271. Solids dissolved in liquids. This is the most interesting kind of solution and is apt to be regarded as the typical one. In such mixtures the solute is usually either a crystalline substance like common salt, or a glutinous substance, known as a *colloid*, such as gelatine. The solvent may be water, alcohol, ether, benzine, or any one of a variety of liquids commonly used to dissolve different substances. Indeed, all liquids are to some extent solvents of something, though there is no "universal solvent" for everything.

Water dissolves a great variety of solids, forming what are known as aqueous solutions. The amount of a solid that can be dissolved by a given quantity of water increases with the temperature in all but a few cases, such as lime, which is more soluble in cold than in warm water.

When an excess of the solid is added to a saturated solution it remains in the solid state; therefore, to insure saturation there should always be an excess of the solute. If a saturated solution of the more usual type is warmed, it ceases to be saturated unless there is an excess of the solute, because at this higher temperature the solubility is higher, and more of the solid is needed to saturate it. If, however, it is cooled, the solubility decreases, and some of the solute reappears in the solid state. This also occurs when the liquid evaporates, and the process may be carried on until nothing is left but the solute, a procedure known as "evaporating the solution to dryness." If the solute is a crystalline substance like salt, it forms solid crystals as a result of either evaporation or cooling, except in those cases to be explained later where the solution becomes supersaturated.

272. Heat of solution. When a solid dissolves in a liquid, work is done upon it in changing its structure from the solid to the more mobile liquid state. The energy required for this transformation is withdrawn from the solution, whose temperature is consequently lowered in the process. The heat corresponding to this energy taken from the solvent by one gram of the solute is known as **the heat of solution**. A consequence of this fact is the usual increased solubility with higher temperature. However, there are substances whose solubility decreases with rising temperature, and which evolve heat

during solution instead of absorbing it. In some cases extremely large heats of solution are observed, but the solutions so formed are not simply solutions but chemical or quasi-chemical unions whose constituents combine with a large evolution of heat in the process of forming a more or less stable compound. This is true, for instance, when lime and water combine, forming a hydrate, with a vigorous evolution of heat. The same is true of phosphorous pentoxide, which is a very powerful absorber of water and evolves much heat with the formation of phosphoric acid. Substances like these, which combine vigorously with water, are called *hygroscopic*, and are very useful as drying agents in connection with air pumps and other apparatus where all moisture must be eliminated.

273. Distillation. Liquids in solution may be separated from the solvent by a similar process. If one of the liquids *A* of the mixture

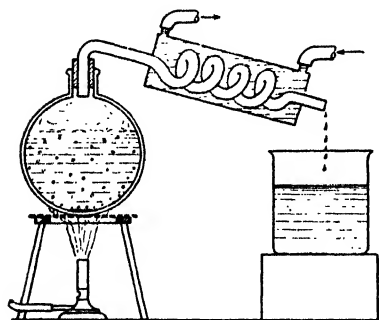


Fig. 35.

has a higher vapor tension than the other component *B*, its boiling point is lower and the solution ordinarily boils at a temperature somewhere between the two boiling points. The resulting vapor contains both liquids, but in general is richer in *A*, the more volatile component. If the vapor is then condensed by coming in contact with a surface kept cool by running water, as shown in Fig. 35, the liquid **distillate**, as it is

called, is also richer in *A*. But as boiling continues, the solution becomes more and more concentrated in *B*, while the boiling point steadily rises and the vapor itself becomes increasingly rich in this less volatile component. Therefore the process is stopped after a certain fraction has been distilled, depending upon the original proportions of *A* and *B*. The resulting distillate is again distilled to obtain a still higher concentration, and by repeated *fractioning* a distillate very rich in *A* may be obtained.

274. Freezing point of solutions. In general, solutions freeze at a lower temperature than the pure solvent. Sea water does not freeze at 0°C , and the stronger its salt content the lower the temperature must be to form ice. A 10 per cent solution of common salt freezes at about -7°C , while the freezing point of a 20 per cent solution is about -17°C .

Raoult discovered that in the case of dilute solutions of non-electrolytes, the lowering of the freezing point varies as the concentration and depends upon the number of molecules of the solute in a given weight of the solvent without regard to the nature of the molecule. This means that a gram molecule of all substances of this type, such as sugar, lowers the freezing point by the same amount.

But when the solute is an electrolyte, that is, a carrier of electricity by means of the process known as electrolysis, its molecules tend to split up or "dissociate" into ions, or carriers of electricity, when in solution. The proportion of these ionized molecules to the total number of dissolved molecules increases with the dilution, so that with infinite dilution they are all split up. There are then n times as many dissolved particles as there would be in the case of a nonelectrolyte similarly diluted, where n is the number of ions formed from a single molecule. Therefore, in the case of electrolytes, the lowering of the freezing point is always considerably greater than that predicted by a literal interpretation of Raoult's law.

We may now restate Raoult's law by saying that *the same number of dissolved particles, whether they are un-ionized molecules, ions, or associated groups of molecules, produce the same lowering of the freezing point in a given quantity of a given solvent.*

275. Cryohydrates. The typical behavior of a salt and water solution is shown by the diagram of Fig. 36. The curve AP gives the freezing point for various concentrations of sodium chloride, beginning with pure water, and ending with a 22.4 per cent solution at P . Similarly BP shows the temperatures and concentrations of saturated solutions when crystallization is about to begin if the temperature is lowered or water evaporated. These curves intersect at the point P , known as the **eutectic point**, below which, along the line PE , both components appear, forming a sort of bisque of ice and salt known as a **cryohydrate**. To obtain the eutectic point we have only to cool a solution of a strength and temperature as indicated by the point C . At D it freezes, but the ice that forms is free from salt, so what remains unfrozen is more

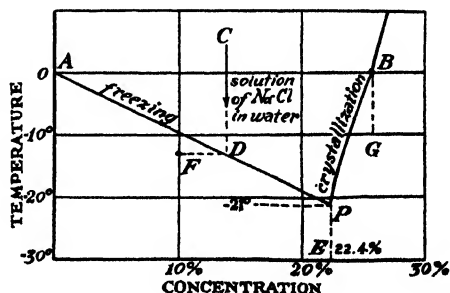


Fig. 36.

concentrated than before. Further cooling carries the solution along the AP curve with steadily increasing concentration, until at P the residual unfrozen liquid solidifies to form the cryohydrate. This has a concentration of 22.42 per cent and forms at -21.2°C .

With the exception of the line PE , the space below the two curves represents an impossible, or unstable situation. For example, a 10 per cent solution of ice and this salt cannot be brought to -13° , as at the point F , without freezing out some of the water, and thus changing the concentration to that of D , nor can a condition like G exist in equilibrium, because a 26 per cent concentration is super-saturated at -10° . As a matter of fact, such a condition may be realized with some substances by cooling a saturated solution, but it is unstable, and a minute crystal of the salt introduced into the mixture results in violent crystallization with an evolution of heat, carrying the temperature up to the BP curve.

A second eutectic point at B is due to a change in the type of crystal when the temperature is 0.15°C and the concentration 26.34 per cent. Some solutions have several such points, but there is always one corresponding to a particular concentration which gives the lowest possible freezing point of the combination. This principle applies also to alloys, which in general melt at a lower temperature than the solvent metal. Solder is an alloy of lead and tin in such proportions that it has a lower melting point than either constituent. A fusible metal, invented by Lipowitz, consisting of 50 per cent bismuth, 27 per cent lead, 13 per cent tin, and 10 per cent cadmium, melts at 60°C , which is far below the melting point of any of its constituent metals.

276. Freezing mixtures. As we have seen, many solid substances (but not liquids or gases) produce a cooling of the mixture during the process of solution. This effect might be used to lower the temperature of the surroundings, but a much more effective type of freezing mixture is one in which one constituent is changed from the solid to the liquid state during the process, when the absorption of the latent heat of fusion by the solution results in a very rapid lowering of the temperature.

Common salt cools water hardly at all when dissolved, but if mixed with ice, the temperature may easily be lowered to -18°C , corresponding to 0°F .† This is because the ice, when the air is above the freezing point, is always moist. This film of water dissolves some

† Especial precautions are needed in order to reach the ideal minimum of -21.2°C .

salt, and the resulting brine melts more of the ice. For every gram melted, 80 calories are withdrawn from the mixture, and the temperature steadily falls as long as both ice and salt remain in the solid form, or until the cryohydrate state is reached.

Crystalline calcium chloride is even more effective than salt in producing a low temperature, for if 100 parts are mixed with 70 parts of snow at 0°C , the resulting temperature is -54.9°C when all the snow is melted. This is due partly to the absorption of the heat of fusion of the ice, and partly to the withdrawal of the heat of solution of the salt.

277. Boiling point of solutions. Salts and many other solids, when dissolved, raise the boiling point of the solution. This is because they reduce its vapor tension, and it must be heated above the boiling point of the pure solvent before the pressure of the vapor is equal to that of the atmosphere. This is shown in Fig. 37, where the solid curve is the steam line of pure water showing its boiling point at 100° under atmospheric pressure. The dotted line is the vapor-pressure curve after a salt has been added. At any given temperature the vapor pressure of the solution is obviously lower than that of the solvent, and it must be heated in order to boil. The diagram shows this condition under atmospheric pressure, when the temperature must rise from 100° at *A* to some higher value at *B*, before the vapor pressure of the solution equals that of the atmosphere so that boiling can occur. At *C*, although the temperature is 100° , the pressure is less than one atmosphere, and only slow evaporation can take place. Raoult found laws for this phenomenon similar to those of the lowering of the freezing point, and the lowering of the vapor pressure in dilute solutions of nonelectrolytes is proportional to the concentration. In electrolytic solutions there is an increased elevation of the boiling point due to dissociation.

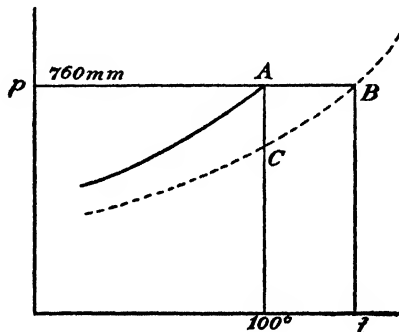


Fig. 37.

278. Diffusion of gases. The process by which two fluids, either liquid or gaseous, tend to interpenetrate each other by the motion of their molecules is called **diffusion**. If, for instance, two jars containing different gases have their mouths in contact with each other, as

shown in Fig. 38, the two gases rapidly mix until each jar contains both gases in the same proportions as the other. This mixing occurs regardless of the relative densities of the gases. If *B* contains the heavy gas carbon dioxide, and *A* is filled with a much lighter gas like



Fig. 38.

hydrogen, hydrogen descends, and the CO_2 ascends until each is distributed evenly throughout the total volume, exactly as if the other were not present, exerting its partial pressure in accordance with Dalton's law.

The only appreciable effect of the pressure of the second gas is in the *rate* at which the first occupies the total volume. If one of the jars has been exhausted, the gas from the other would fill it with explosive violence, whereas the process of mixing just described may require many minutes if the volumes are fairly large. The obvious reason for this is that collisions are constantly occurring between the molecules, which limit their mean free path and so retard the process of diffusion.

279. Diffusion through porous solids. If the gases referred to in the last paragraph had been separated by a sheet of paper, or any other porous body, the interpenetration would still have taken place, though at a slower rate. Also the gas of lower density would diffuse through the porous sheet more rapidly than the other. This is because the flow, or *effusion*, through the minute holes of the separating medium varies inversely as the square root of the density of the gas, as was proved in Article 158. Therefore if two gases are separated in this way, the ratio of velocities u_1 and u_2 with which they pass into each other's domain is given by $u_1/u_2 = \sqrt{d_2/d_1}$, where d is the density.

This results in unequal mixtures at first, as may be shown by the following experiment: A porous earthenware jar *A* is fitted tightly with a stopper through which passes the bent glass tube of narrow bore *T*, shown in Fig. 39. This tube is partly filled with water, colored to make it easily seen, and standing normally at the level *aa*. An inverted glass beaker *B* is supported over the jar, and a pipe *P* attached to a gas-cock enters *B* from below. At first *A* and *B* con-

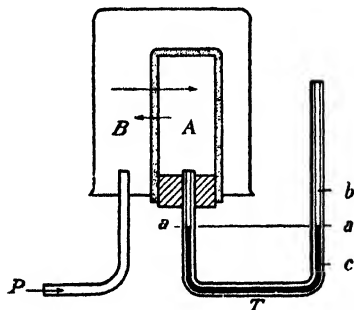


Fig. 39.

tain only air at atmospheric pressure. Then the gas is turned on for a few seconds, during which it rises because of its low density into *B*, temporarily driving out most of the air, and at once begins diffusing into the space *A*, while the air in *A* also diffuses into the gas contained in *B*. But the two rates of diffusion through the porous walls are not equal. The lighter illuminating gas passes more rapidly through the pores than the heavier air, so that for a short time the pressure in *A* is raised by an addition of the lighter gas, without a corresponding loss due to the slower outward flow of the heavier air. This is indicated on the diagram by the two arrows of different lengths, and the result is an increased pressure in *A* which drives the liquid in the glass tube up to a new level *b*, often several inches above *a*. This unbalanced condition lasts for only a short time, because, as we have seen, the ultimate proportions of the two gases tend to become the same in both regions, under atmospheric pressure. While this steady state is being realized, the pressure in *A* falls, and the liquid in the manometer returns to *aa*.

Now if *B* is removed, a reverse process takes place. The porous jar contains a mixture of illuminating gas and air, but is surrounded by air only. The gas starts to diffuse outward more rapidly than air can enter to take its place. This results in a temporary fall of pressure in *A*, and the manometer level sinks to *c*, but ultimately returns to *a* when the inward diffusion of air has again restored atmospheric pressure.

280. Diffusion of liquids. If two liquids which are soluble in each other are placed in contact, with the heavier one below, a slow process of diffusion takes place which ultimately results in a homogeneous mixture of both. In principle this is identical with the diffusion of gases, but takes place much more slowly, owing to the much smaller mobility of the molecules. In fact it requires weeks or even months to diffuse completely, depending upon the nature of the liquids, their concentration, and the form of the containing vessel. A vertical tube containing water with a concentrated solution of copper sulphate at its lower end illustrates the diffusion of a salt. Its progress is indicated by a gradual rising of the blue color, and if the length of the tube is, say, ten times its diameter, many months elapse before the coloring becomes sensibly uniform.

The common acids and solutions of crystalline bodies, such as mineral salts, diffuse much more rapidly than the gummy substances known as colloids, and a mixture of the two types may be partially separated by making use of the different rates with which they diffuse

through a porous diaphragm. The molecules of the dissolved salt pass readily through the minute holes of some membranes such as parchment paper, while the much larger colloidal molecules are almost wholly stopped. The membrane thus acts like a sieve in separating colloids from crystalloids, and is much used in chemical analysis. The process is called **dialysis**.

281. Osmosis. The diffusion of liquids, being due to the same cause as that of gases, namely, the motion of their constituent molecules, must be expected to follow laws similar to those of gases. This has been found to be the case, though we must adapt our ideas regarding pressure, density, and so forth, to meet the conditions imposed by the solution of solids or liquids in liquids, instead of the more familiar and simpler conditions of gaseous mixtures.

The pressure with which a solute tends to diffuse throughout a solvent is known as **osmotic pressure**, and the process by which this

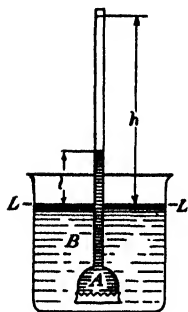


Fig. 40.

pressure is made evident is called **osmosis**. Certain membranes, like bladder, have the property of allowing free passage to water, while they greatly retard the molecules of a salt dissolved in it. These are called semipermeable membranes. If such a membrane separates distilled water from a salt solution, the tendency to mix in equal proportions on either side takes place mainly in one direction, that is, by means of water passing into the solution and diluting it progressively until, in theory, it reaches infinite dilution and so becomes pure water also. In doing this the dissolved salt occupies a larger and

larger volume just as if it were a gas free to expand, and exerts a pressure which may be measured as follows: A thistle tube *A* (Fig. 40) has goldbeater's skin tightly stretched over its mouth, and is filled to the level *L* with a concentrated solution of some salt like copper nitrate. It is then immersed in distilled water *B* as shown. The salt cannot readily get through into the water, but it is diffused none the less by the water that passes through the membrane and causes a gradual rise of the column in the stem. At some level *l* this process ceases, because the solution is steadily growing more dilute, and the hydrostatic pressure due to *l* ultimately balances the osmotic pressure, which decreases with increasing dilution. If the solution could be kept at its original strength, the final height *h* would measure the osmotic pressure of a concentrated solution according to the familiar equation $p = hdg$.

The apparatus just described is not very satisfactory except with very dilute solutions, because the membrane is so fragile that it is likely to break under the weight of a long column above it. Pfeffer, a German botanist, in 1877 succeeded in making very tough membranes by depositing copper ferrocyanide as a film within the walls of porous jars. In this way he measured pressures as high as 307.5 mm of mercury, using a 6 per cent solution of sugar. As a result of these investigations, Pfeffer concluded that osmotic pressures varied as the concentration, provided the solution was sufficiently dilute. This is true for nonelectrolytes, but when dissociation takes place there is an abnormal increase in the osmotic pressure, just as in the case of the depression of the freezing point, and the elevation of the boiling point.

In 1887 van't Hoff showed that Pfeffer's conclusion was equivalent to Boyle's law for gases, since concentration varies inversely as the volume occupied by unit mass of the solute; therefore $p \propto 1/v$, or pv is constant at constant temperature. He also showed that osmotic pressure varies as the absolute temperature, which is equivalent to Charles' law, and that osmotic pressures of different substances, not dissociated, when present in amounts proportional to their molecular weights, are all equal, which is analogous to Avogadro's law.

Thus dissolved substances behave in many respects like gases. They have the same gas constant R per mole in dilute solutions, and exert the same pressures, as would be expected of the same number of gas molecules in the same space at the same temperature, while their state is determined by the same equation $pv = rT$.

Osmosis plays a very important part in both animal and vegetable physiology. This was first studied by DeVries who found that many cell walls are semipermeable and therefore enlarge or contract according to whether they contain a less or more concentrated salt solution than their surroundings. Cysters behave in this way, and swell when placed in fresh or brackish water, because when first taken from the ocean they contain salt water, and in less saline surroundings, fresh water passes through their membranes to dilute their more concentrated contents, thus making them appear fatter than before.

SUPPLEMENTARY READING

- S. Arrhenius, *Theory of Solutions*, Yale University Press, 1912.
A. T. Lincoln, *Textbook of Physical Chemistry* (Chapters 14, 15, 23), Heath, 1918.

CHAPTER 22

Propagation of Heat

282. Modes of propagation. There are three ways in which heat may be carried from one place to another. These are convection, conduction, and radiation. Convection means literally *carrying with*, and involves bodily transfer of heated substances, such as the *feuerwerfer* (fire-thrower) used in the World War, and the "Greek fire" of the ancients. Conduction (*leading with, or along*) is the slow process by which heat is carried through a substance by molecular activity, the molecules of the hotter portions giving up some of their kinetic energy to adjacent and less active molecules in the colder portions. Radiation is a transfer of heat in the same manner as light, and with the same velocity. The mechanism of this process is not fully understood, though wave motion plays an important part in it, as well as a quasi-corpuseular transmission of bundles of energy known as *quanta*.

283. Convection in nature. This mode of heat transfer is the most obvious, and, except for the heat we get from the sun, the commonest in nature. The heat of the sun in equatorial regions warms the air, which rises. This causes the trade winds, which blow from higher and cooler latitudes toward the low-pressure region produced by the expansion of the air in the tropics. But the heated air carries its warmth far north and south of the equator, and thus by convection helps maintain the equable climate of the temperate zones.

Ocean currents, like the Gulf Stream, are notable examples of convection. Whatever is their cause, they carry great masses of warm water from the tropical oceans into temperate and even arctic latitudes, thus greatly modifying the climate of the countries they approach. These warm streams are compensated by cold arctic currents like the Humboldt Current, which flows from the Antarctic Ocean up along the western coast of the United States, making the coast of southern California cooler than would be expected from its latitude.

284. Convection in boilers. Before a pot of water boils, convection currents are in constant circulation, carrying the hottest water

from the lower surface in contact with the fire up to the top, while water from the cooler upper layers descends to take its place. These currents are easily seen if the water contains small solid particles held in partial suspension, and they become increasingly marked as the boiling point is approached. The convection currents in boiling water are extremely vigorous.

The boiler of a steam engine has to be carefully designed to allow for this kind of circulation, for if it did not occur, the layers nearest the grate would become superheated and form steam with explosive violence, blowing the cooler water far above its normal level and even wrecking the boiler. It is especially important to consider the circulation of convection currents in the design of water-tube boilers, where much of the liquid is contained in iron pipes of relatively small section. Such an arrangement is shown in Fig. 41, where the hot gases from the grate *G* pass upward through the system of water tubes *T*, and out by the back connection *B* to the chimney. The gases pass-

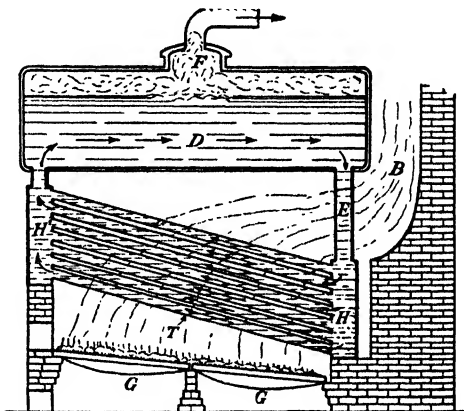


Fig. 41.

ing under the drum *D* and past the tube *E* have already lost some heat, so that the water at the back of the boiler is cooler and denser than that in the front portion, and descends as indicated by the arrow through the back header *H* where it starts upward through *T*, meeting progressively hotter gases as it ascends. This ascending current grows steadily hotter and less dense until it reaches the front header *H'*, from which it again enters the drum. The most vigorous boiling occurs under the steam dome *F*, where the pressure is always a little lower than elsewhere because of the outward flow of steam which occurs at that point by way of the steam pipe leading to the engine.

285. Central heating. Houses heated by a furnace depend upon convection for the transfer of heat from the furnace to the air of the rooms. In hot-air furnaces the cold air from outdoors is warmed by contact with the hot iron of a box or other container heated by the fire. It expands, rises, and flows through pipes of large section with

as few bends as possible into the various rooms to be heated. The colder air from these rooms escapes through chimney, door, or window. But if no escape besides stray leakage is possible, the system is very ineffective.

Hot-water furnaces employ a similar system using water as a conveyor of heat instead of air. The water is heated in a boiler *A* (Fig. 42) and flows upward through pipes to "radiators" *B*. These warm the air near them, causing it to rise and give place to cooler air which is itself warmed in turn. The water, having thus lost heat, now becomes denser and flows down through another pipe to the boiler to be heated over again. An expansion tank *C* connected to the hot-water system is a necessary part of the equipment. It allows for changes in volume of the circulating water and so prevents the pipes from bursting as the water expands.

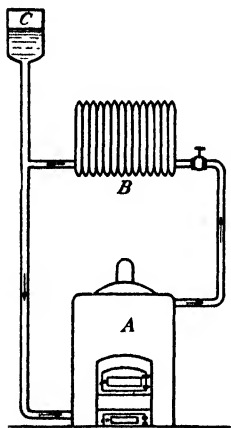


Fig. 42.

In steam-heating systems the water in the boiler actually boils, and the resulting steam, but not the water, is carried through the radiators, where it condenses in contact with the relatively cool metal, gives up its heat of vaporization, and runs down again as water to the boiler. This is a very efficient method, for when steam condenses it gives up just as many calories as were put into it when it was formed from water at the boiling point. However, steam heating suffers from the disadvantage that no heat is delivered until the water boils; consequently the radiator in condensing the steam is also generally raised to nearly 100°C , and lower temperatures for moderate heating are less easily obtained than with the hot-air or hot-water systems.

286. Prevention of convection. In some cases, convection of heat is undesirable. We wear clothes to prevent the heat of the body from being conveyed away by air currents. Wool is particularly effective in this way because its kinky fibers retain a layer of motionless air near the skin, thus preventing convection currents, and also reducing conduction losses to a minimum, as we shall see later.

Everyone knows that he chills much faster when the wind blows than when the air is still. This rapid removal of heat, due to convection currents set up even in woolen clothes, may be counteracted by wearing leather or closely woven canvas outside the woolen garments.

So-called "fireless cookers" work on a similar principle. The food is first partly cooked over a fire, and is then placed inside a container surrounded by straw, felt, or still better, a vacuum, so that little heat escapes and slow cooking continues for several hours. Thermos bottles function in the same way. They are made of double-walled glass with most of the air between the walls exhausted. Thus heat cannot pass in or out either by convection or conduction, while radiation is minimized by silvering the outer layer, according to a principle to be explained later.

Incandescent lamps are partly exhausted to prevent the air from carrying heat away from the filament and to prevent its oxidation. If this were not done it could still be heated to incandescence by the current, but more electrical energy would be required to supply the increased loss, and its life would be shortened by oxidation.

287. Conduction. This molecular process is very slow compared to the usual cases of convection. A silver spoon dipped into boiling water may be held in the hand for several minutes before becoming uncomfortably hot, while a stick of wood under the same circumstances gets warm so slowly that one does not observe any rise of temperature. Evidently the rate of conduction is different with different substances.

The reason why stone or metals feel hotter to the touch on a hot day, and colder on a cold day, than cork, wood, or cloth, is that stone and metal conduct heat to the skin, or carry it off more rapidly at a given temperature, than other substances of lower conductivity. Thus common experience teaches us that some bodies, especially metals, are good conductors of heat, while others, especially those of animal or vegetable origin, are poor conductors.

Among the poor conductors are gases and liquids. It is difficult to prove this fact experimentally, because convection currents are generally set up unless great care is taken to prevent them, and the heat carried in this way masks their insulating property. However, the layer of motionless air held in the packing of a fireless cooker is an evidence of its insulating power. To prove that liquids act in the same way, we may heat the upper surface instead of the bottom of a liquid contained in a vessel, and thus eliminate convection currents. In this case the lower layers remain cold, even when an inflammable liquid like mineral oil is burning on the top of a beaker of water.

288. Coefficient of conductivity. Experiment shows that when one portion of a bar of uniform section is heated, the total transfer of heat along it varies as the area of the section, as the time during which

it flows, as the temperature gradient, and as the nature of the substance. This may be expressed by the following equation, which gives the amount of heat transferred past a given section of the bar in a given time:

$$H = ka\left(\frac{t_1 - t_2}{l}\right)\tau, \quad (1)$$

where $(t_1 - t_2)/l$ is the space rate of change of temperature, or gradient, assumed constant for a distance l , a is the cross section, and τ is the time, while k is a constant known as the coefficient of conductivity of the material in question. If the bar is a cube having one square centimeter cross section and is one centimeter long, and if a difference of one degree centigrade is maintained between opposite faces, then in one second the number of calories which flow through it is numerically equal to k , because everything else in the right-hand member of the equation is unity. The preceding definition assumes that there is no loss of heat from the sides of the conducting bar. If there

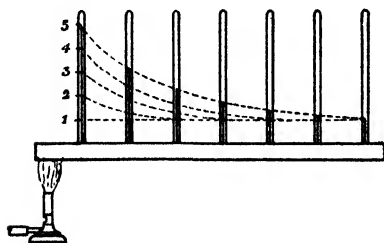


Fig. 43.

is such a loss, the temperature gradient is variable, becoming less and less steep at increasing distances from the source of heat. At a given point it is then expressed as dt/dl .

The flow of heat along a bar, with steady loss from all surfaces, may be studied by placing thermometers in small holes drilled at

regular intervals along it, as shown in Fig. 43. The line numbered 1 shows the initial condition when the thermometers are all at the same temperature. After heating the left-hand end for a short time, the temperature is shown by 2, while 3 and 4 illustrate conditions at later time intervals. Where the slope is steepest the heat is flowing most rapidly, as is to be expected from the equation (1).

289. Measurement of conductivity of solids. In order to measure the conductivity, k , of a solid, it is necessary to know all the other quantities involved in its defining equation. This may be accomplished by the method devised by Searle, shown in Fig. 44. A short thick bar of the metal to be studied is packed in felt to minimize loss by convection or conduction from its surface. The relatively large cross section also contributes to this end, because the rate of flow depends upon a , which varies as the square of the diameter d , while the surface per unit length varies only as the first power of d . There-

fore, when d is made large, the effect of surface losses becomes less and less important in comparison with the amount of heat conducted through a .

A steam chamber at one end is supplied with steam from a boiler not shown, while a coil through which water runs keeps the other end cool. Four thermometers are placed as indicated, two to measure

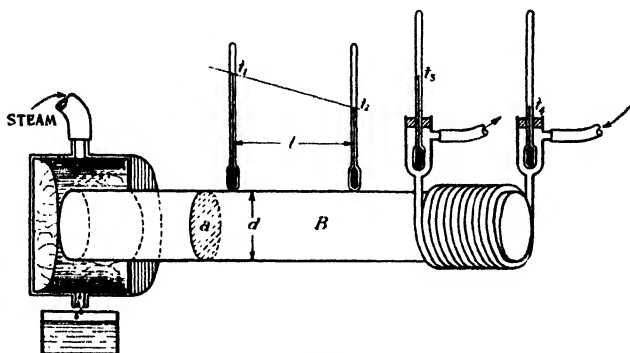


Fig. 44.

the temperature gradient, and two more to measure the temperatures of the incoming and outgoing water. It takes some time to reach a steady state, or *regime*, after turning on the steam, but in time all the thermometers become almost steady, provided the steam and water flow at nearly uniform rates, and the latter enters at a constant temperature t_4 . The temperature gradient is now a straight line as indicated on the diagram, and is readily obtained from t_1 , t_2 , and the distance l . The amount of heat which flows through during the time τ is obtained from the measured amount of water delivered from the coil during this time multiplied by its rise in temperature $t_3 - t_4$. The section a is readily measured, so that k may be calculated.

Substance	k	Substance	k
Aluminum.....	0.504	German silver.....	0.07
Copper.....	0.918	Glass.....	0.0015
Iron (pure).....	0.161	Oak wood.....	0.0006
Lead.....	0.083	Pine wood.....	0.0004
Mercury.....	0.0197	Paper.....	0.0003
Silver.....	0.974	Porcelain.....	0.0025
Zinc.....	0.265	Rubber.....	0.00045
Brass.....	0.260	Ice.....	0.005

As is seen from the conductivity equation, the dimensions of k are those of H divided by time, temperature, and the first power of a length, since a involves length squared while l is the first power only. Values of conductivity are then expressed in calories per centimeter per second per degree centigrade. On page 255 are given values of k for a few of the more important metals, alloys, and other materials at an average temperature of about 18°C .

290. Measurement of conductivity of liquids and gases. Liquids may be heated in a manner similar in principle to that described above, but the heat should be applied at the upper surface of a shallow layer. Under these circumstances, convection currents are not set up to any appreciable extent. The heat is best supplied electrically, because of ease in regulation, and in calculating the amount generated from the measured current and resistance of the heating coil. Solids of such low conductivity that thin sheets must be used instead of bars, are examined in much the same way. The sheet or slab is placed between the electric heater and a metal plate, whose temperatures are measured after a steady state has been attained.

It is much more difficult to measure the thermal conductivity of gases, for convection currents are unavoidable, and radiation also takes place across the space occupied by the gas. The general method is to measure the loss of heat by a body warmer than its surroundings and suspended in a vessel filled with the gas to be studied, which is reduced to a pressure low enough to eliminate convection currents. If it is not too low, both experiment and theory show that the conductivity is not affected by the pressure.

Loss of heat due to convection currents is found to be negligible if the pressure is below 15 cm of mercury, so that at, say, 5 cm pressure, the cooling of the heated body is due only to conduction and radiation. The rate of cooling at this pressure is observed, and again when the space has been exhausted as completely as possible. In this condition the cooling is due to radiation alone. So having eliminated convection, and measured the radiation loss, we can calculate the conductivity.

Conductivities for certain liquids at about 20°C and gases at about 0°C are given in the table on page 257. It will be seen that the order of magnitude of the latter is considerably smaller than that of the former, while in general, among gases, those of low molecular weight are the best conductors. Thus carbon monoxide is a better conductor than carbon dioxide, and nitric oxide (NO) is better than nitrous oxide (N_2O).

Liquid	k	Gas	k
Alcohol.....	0.00043	Hydrogen.....	0.00032
Glycerine.....	0.00058	Helium.....	0.00034
Ether.....	0.0003	Oxygen.....	0.000056
Mercury.....	0.0152	Air.....	0.000052
Paraffin Oil.....	0.00035	CO.....	0.00005
Water.....	0.00143	CO ₂	0.00003

291. Rate of temperature rise. A familiar type of experiment used to illustrate comparative conductivity consists in coating rods of several metals with a paint that changes color when hot.

The rods, held vertically, are heated simultaneously at their lower ends by boiling water, and the rate at which the exposed portions change color gives a rough index of their relative conductivity. This experiment, however, is somewhat misleading, because the rate at which the temperature rises along such a bar is not the same thing as the rate at which heat crosses a given section. The latter depends upon the value of k for a given temperature gradient and section, while the former involves in addition both the specific heat and density of the metal. Thus a substance of low specific heat would grow warm at the far end faster than another metal of high specific heat but having similar density and conductivity. This is because it takes less time for a given amount of the metal of low specific heat to acquire a stated temperature, when supplied with heat at the same rate, than a metal of higher specific heat.

292. Radiation. Our most important illustration of this mode of heat transfer is the radiation of heat from the sun. It comes to us in exactly the same way as the sun's light, and with the same velocity of 3×10^{10} cm/sec., or 186,000 mi./sec. It is the chief source of terrestrial heat, because the earth's surface receives very little heat from the interior, and the amount we owe to the stars is negligibly small. Such apparently different phenomena as radio broadcasting, radiant heat, light, X-rays, the gamma rays from radium, and possibly the "cosmic rays" from outer space, are all essentially the same, but differ from each other in wave length. "Radio" waves are the longest, those of radiant heat come next, and so on in the order named to the excessively short waves emitted by radium and other radioactive substances and the still shorter cosmic rays, if they are of a wave nature.

These various forms of radiation travel best through empty space, but all can penetrate matter, even solids, with more or less freedom.

However, the laws describing their transmission, absorption, and other characteristic properties will be discussed later under the sections of light and electricity. At present we need think of radiant heat only as a stream of energy flowing out from the radiating body with incredible speed, and stopped only when absorbed or reflected by matter whether in the gaseous, liquid, or solid form.

293. Inverse square laws. When a body radiates heat, it loses energy. This radiant energy tends to spread out in all directions, as the very word *radiation* implies, but its total amount remains unchanged until it meets some material substance which absorbs a portion of it and in so doing reconverts this portion into heat. This is much the same as the absorption of the long waves of radio transmission by the antennae of a receiving station, where they are reconverted into electric currents similar to those that originally caused the radiation.

If the radiator is a point or sphere, and surrounded by empty space or a homogeneous medium that is only slightly absorbent, like air, the energy spreads out equally in all directions, so that its intensity grows steadily less as the distance from the source increases. Intensity I is defined as the amount of energy which passes each second through a plane of unit area (one square centimeter) perpendicular to the direction of flow. Let the total energy radiated per second from a point O be H ergs/sec. Then circumscribe

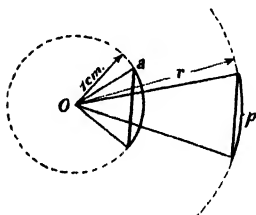


Fig. 45.

a sphere of unit radius about the point, as shown in Fig. 45. Since the area of such a unit sphere is $4\pi \text{ cm}^2$, the rate of flow across the unit area a is given in ergs per second by $I = H/4\pi$. This is the flow through a unit solid angle, and if $H = 4\pi$ ergs/sec., the intensity over any part of the sphere is unity.

Now consider a point p distant r cm from O . The area of a sphere passing through p and centered on O is $4\pi r^2$; therefore the rate of flow across unit area (intensity) at p is $H/4\pi r^2 = I_p$. If we compare I_p with I above,

$$\frac{I_p}{I} = \frac{H}{4\pi r^2} \times \frac{4\pi}{H} = \frac{1}{r^2}.$$

This means that *the intensity varies inversely as the square of the distance from the source*, provided there is no absorption. This is known as the inverse square law, and it enables us to compare the intensities at different distances from a common source. Thus if r_1 and r_2

are the distances, since $I_1 = I/r_1^2$ and $I_2 = I/r_2^2$, it follows that $I_1/I_2 = r_2^2/r_1^2$.

There is nothing in this law peculiar to heat or light for it is a geometrical property of three-dimensional space that areas of similar solids increase as the square of their linear dimensions. Hence anything which goes out from a point uniformly in all directions, spreads itself over areas which increase as the square of the distance from that point, so its intensity is diminished at the same rate. The spines of a sea urchin, assuming uniform radiation, would obey this law, so there is nothing mysterious about it unless we regard three-dimensional space as a mystery. If our space had four dimensions, radiant energy would fall off inversely as the cube of the distance, while in a plane it falls off as the first power, a fact observable in the expanding wave front which radiates from a stone thrown into a pond. But if the propagation of energy were confined to a straight line, and no absorption occurred, the intensity would remain constant to infinity.

294. Measurement of thermal radiation. The most obvious way of measuring radiant heat is by absorbing it in some substance whose heat capacity is known, and then observing the rate at which its temperature rises. The problem is then purely calorimetric, and H is found from the usual relation $H = (\Sigma sm)(t_2 - t_1)$. This is made use of in practice for measuring the rate at which solar energy reaches the earth. The apparatus is essentially a calorimeter containing water which is heated through a thin diaphragm of blackened metal. Thus the apparatus almost completely absorbs the radiant energy which falls upon it. If due corrections are made for the absorption of the atmosphere, the earth, when the sun is vertical, receives on an average 1.937 calories per square centimeter per minute, or about 1.8 horsepower per square meter. This figure S is known as the **solar constant**, although it is not constant, but varies with the sun spots over an eleven-year cycle, and must be gradually growing less as the sun loses energy.

A much more delicate device for detecting and measuring radiant energy was invented by Sir William Crookes, and is called the radiometer. Its essential features are shown in Fig. 46, where A and B are two mica vanes at the ends of a light aluminum bar supported by a quartz fiber in a highly exhausted vessel. If each vane is coated with lampblack on one face and silvered on the other, the

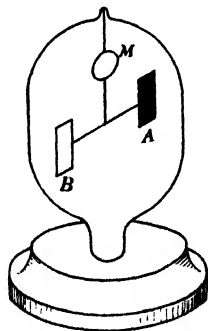


Fig. 46.

blackened surface will absorb the radiant heat that falls upon it and so grow warmer, while the polished face will reflect almost all it receives and remain cold. If now a beam of radiant heat falls upon the system, or, still better, is directed at the blackened face of *A* alone, the latter becomes warmer, and gas molecules striking it rebound with greater energy than those which strike its back or either face of *B*. This results in an unbalanced force tending to push *A* backward, and the resulting torque rotates the system, which carries a small mirror *M*. Then if a beam of light falling upon *M* is reflected to a scale, very minute angular displacements may be read and measured from the displacement of the spot of light. The degree of exhaustion is a most important consideration in this instrument. If it is not good enough, collisions between the molecules themselves, after impact with the black vane, result in a general rise of temperature of the gas, which neutralizes the effect desired, while if the pressure is too low a new effect, to be described later, appears and reverses the rotation.

The whirling radiometers seen in opticians' windows work on this plan but are not adapted to quantitative measurement. The bolometer, thermopile, and radiomicrometer are also very sensitive detectors of radiant heat, but as they are electrical in principle, an explanation of their behavior must be postponed.

295. Prévost's theory of exchanges. This fundamental principle was first formulated in 1792 by Pierre Prévost, a Swiss philosopher, man of letters, and physicist. Among other important contributions to the theory of radiation, he showed that all bodies are continually radiating heat, and at the same time absorbing it from their surroundings. Thus there is set up a process of "exchanges" of energy, resulting in warming objects that receive more than they give, while those that radiate more than they receive must lose heat. In the former case the temperature of the body rises, while in the latter case it falls. If we combine this principle with the second law of thermodynamics, it is evident that hotter bodies, in the presence of cooler ones, tend to grow cooler by radiation alone, while the latter tend to grow warmer. This may be strikingly demonstrated by placing a block of ice in front of a whirling radiometer, with the result that the sense of rotation is reversed long before the highly exhausted gas in the bulb can have lost appreciable heat by conduction or convection. The ice, though radiating heat, receives much more from the radiometer (and other surroundings) than it sends out. The blackened surfaces of the vanes radiate more vigorously than the polished ones, without receiving compensation from the ice. They are cooled in consequence, and the

rebounding gas molecules are slowed down by impact with the cooling surface.

If a body and its surroundings are at the same temperature, then the exchange must be equal, for it is a matter of common experience that their temperatures remain constant in accordance with the second law of thermodynamics.

296. Absorption, transmission, and reflection. The ability of bodies to absorb or emit, transmit or reflect, radiant heat depends upon their physical properties. Some, like charcoal, absorb almost all they receive. Others, like rock salt, transmit with great freedom, while still others, like polished metal, do little of either, but reflect radiant heat almost undiminished.

The **absorptivity** of a medium or substance is the ratio of the heat absorbed per second over a square centimeter by a thickness of one centimeter, to the total heat received in the same time by the same area. If these rates are denoted by h_a and h , respectively, the absorptivity is given by $a = h_a/h$. But if there is no transmission, h_a/h represents a purely surface absorption, and then a measures surface absorptivity rather than that of a medium.

Reflectivity is measured in the same way by a coefficient defined by $r = h_r/h$, where h_r is the rate at which energy is reflected from unit area which receives it at the rate h .

Transmissivity, or diathermancy, measures the ability of media to transmit radiant heat even when they are opaque to visible light. It is defined by $d = h_d/h$, where h_d is the rate at which heat is transmitted through a centimeter thickness of the medium of unit area.

Since the heat received by a body per second must be either absorbed, reflected, or transmitted, and since all three actions usually occur to some extent, it is obvious that $h_a + h_r + h_d = h$. Therefore

$$a + r + d = \frac{h_a}{h} + \frac{h_r}{h} + \frac{h_d}{h} = \frac{h}{h} = 1.$$

If there is no transmission, $a + r = 1$; similarly, in the unlikely case of negligible absorption, $r + d = 1$. But the most interesting case is where r may be disregarded. Then $a + d = 1$, a result showing that all the radiant energy not absorbed must be transmitted. In this case, increasing thickness increases the total amount of heat absorbed, while decreasing the amount transmitted. Since d is the proportion transmitted by a centimeter thickness, two centimeters transmit $d \times d$ of the amount received, and n centimeters transmit only d^n of the original energy.

297. Values of transmissivity. This property of substances is profoundly affected by the temperature of the source, so that a numerical value of d is not very significant unless that temperature is specified. The following table gives values of the proportion trans-

Substance	Temp. of Source (°C)	Rel. Trans- mission (%)
Rock Salt.....	390	92
“ “.....	100	92
Plate Glass.....	390	6
“ “.....	100	0
Calcium }.....	390	42
Fluoride }.....	100	33

mitted, as found by Melloni, for layers of substances 2.6 mm thick instead of one centimeter. It is evident from the figures that transparency to visible light is no index of the transmissive power of heat waves, for all the substances named are highly transparent in the usual sense of the word. We see also that with high temperatures

transmissivity increases, except in the case of rock salt. Water is extremely opaque to radiant heat, while alcohol and kerosene are considerably less so, and carbon bisulphide is nearly six times as “transparent” as water. If this fluid is colored with iodine so as to be nearly opaque to visible light, it still transmits so much heat that the beam from an arc light may be brought to a nearly dark focus by a spherical flask containing the solution, and a match may be ignited by holding it there, as indicated in Fig. 47.

298. Emissive power. The ability to radiate heat, called **emissive power**, depends not only on the temperature of a body, but upon its nature, especially that of its surface. It may be defined as the total energy emitted per second per square centimeter of the radiating body, and will be denoted by E' . A comparative idea of the emissive power of various metallic surfaces is easily obtained from the use of “Leslie’s cube.” This a cubical box with an opening at the top by which it is filled with boiling water. The four lateral faces are of different metals differently treated, such as iron painted white, iron painted black, polished brass, and tarnished copper. A thermopile placed at the same distance from each face in turn serves to compare the radiation from these faces. In the case supposed, the polished brass is much the poorest radiator, with tarnished copper considerably better, while the two painted surfaces are almost alike, with

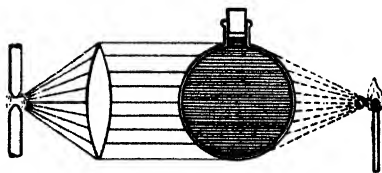


Fig. 47.

the smoother one exhibiting the lower emissive power. This is quite as likely to be the black surface as the white one, because color is of little importance at such a low temperature.

299. Emissivity. The ratio between the emissive power E' of a given surface to that of the perfect radiator E is called the **emissivity** of that surface. This quantity will be denoted by $e = E'/E$, and its value is obviously unity, or 100 per cent, in the case of a perfect radiator, for then $E' = E$.

300. Kirchhoff's law. This law, derived as a result of experiment, states that the ratio of emissive power of a body to its absorptivity, $E':a$, is the same for all bodies at the same temperature, but that its value depends upon the temperature of the source and upon the wave length of the radiation. In other words, $E':a$ is a constant for all bodies at the same temperature and emitting radiations of the same wave length. Hence good absorbers are good radiators, and poor absorbers are poor radiators in exact proportion to their absorptivity.

The absorptivity a differs widely among different substances and differs in the same body with different wave lengths. Lampblack absorbs about 96 per cent of *visible* light, and platinum black 98 per cent. Therefore an ideal body, which absorbs 100 per cent of the energy of all wave lengths it receives, is known as a "black body." This may be realized by using a hollow receptacle, preferably spherical, painted black inside, and with a hole to admit radiations. These, after entering the enclosure, are completely absorbed because of the repeated internal reflections, as indicated in Fig. 48, and only those that enter axially, like the ray r , have any chance of emerging after a very imperfect reflection from the tip of the blackened cone at b . Therefore the absorption is practically perfect, and its coefficient a may be taken as unity.



Fig. 48.

Now, since according to Kirchhoff's law, $E':a$ is a constant, it follows that a black body must be an ideal radiator because it is an ideal absorber. Its emissive power is therefore E by hypothesis, and its absorptivity is unity, or 100 per cent. Then the ratio $E':a$ for any body must equal the same ratio for a black body where $a = 1$. Hence

$$\frac{E'}{a} = \frac{E}{1} \quad (1)$$

This tells us that the numerical value of Kirchhoff's ratio is precisely the emissive power of a black body at the same temperature.

Finally, since the emissivity is defined as E'/E , $E' = eE$, and substituting in the equation above, we obtain

$$\frac{eE}{a} = \frac{E}{1} \quad (2)$$

$$\therefore a = e, \quad (3)$$

or the absorptivity and emissivity of all bodies are equal at the same temperature insofar as they obey Kirchhoff's law. From this it follows that if the hollow sphere shown in Fig. 48 were heated, it would emit more energy per square centimeter through the orifice than any body that was less perfectly "black."

301. Experimental demonstration. An experiment due to Ritchie demonstrates Kirchhoff's law in a convincing manner. A hollow cylindrical drum *A* (Fig. 49) mounted on a glass post, has one of its

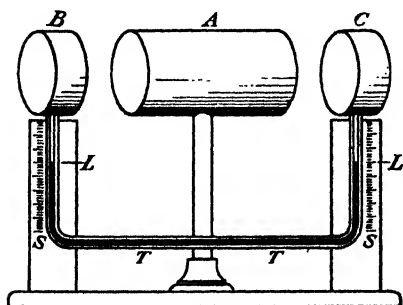


Fig. 49.

end faces of polished silver or nickel, and the other coated with lampblack. When filled with hot water, the former radiates very feebly, while the latter radiates nearly 100 per cent of black-body radiation at the same temperature. The hollow drums *B* and *C* each have one polished and one black face also, and they are connected by a capillary tube *T*, partly filled with a colored liquid

whose two free surfaces stand normally at the common level LL' , indicated by the scales *S*. If the air in either *B* or *C* is warmed, it expands and forces the liquid index down on one side and up on the other. All three cylinders may be rotated about their supports as axes, so that either face may be presented to its opposite neighbor. Then if the blackened face of *B* is opposed to the blackened face of *A*, while two polished surfaces face each other between *A* and *C*, the result is a strong transfer of radiant heat from *A* to *B*, and almost none from *A* to *C*. This results in a falling of *L* and rising of *L'*. But if *bright faces dark*, from *A* to *B*, and *dark faces bright* from *A* to *C*, the levels of the liquid remain at the same height.

The fact that a strongly absorbing surface is therefore a good radiator may also be roughly shown by painting a spot of India ink on a strip of platinum foil. Its black color means that it is absorbing the

light which falls upon it. But if the foil is now heated to redness, the spot of ink becomes brightly incandescent, emitting much more light than the surrounding foil. This indicates a greater emissive power for the short waves of visible light, just as it previously absorbed them more completely at the lower temperature.

302. Applications. Among the many interesting applications of the principles we have been discussing are the following: A thermos bottle has its outer surface silvered to prevent radiation, which would otherwise take place across the vacuum between its glass walls. Radiators in houses are purposely rather rough and unpolished so they may have a high emissivity, and the radiating surface is made as large as possible. A thermometer whose bulb has been coated with lampblack not only records a higher temperature in the sun than one not so treated, but it falls more rapidly to air temperature when placed in the shade. Smooth white duck clothes reflect the radiant heat of the tropical sun more effectually than dark rough woollens, which absorb it instead. The low transmissivity of glass to radiant heat makes it valuable as a fire screen, acting as a shield from the glare of an open fire. For the same reason it cuts off so much of the sun's radiant heat that a "burning glass" does not work so well inside a glass window as out of doors. A "cold frame" for raising early vegetables, and a greenhouse, transmit mostly visible light from the sun through their glass panes, but this light, falling on the ground, is absorbed and warms it. The ground in turn radiates heat, which cannot readily escape because of the poor transmissivity of glass for thermal radiation. Dew forms on the grass on clear nights in summer because of rapid radiation from the heated soil. This radiation causes the temperature of the soil to fall below the dew point, and so condense the moisture in the air. But on cloudy nights the radiation is partly returned by the clouds to the earth, and the cooling process takes place too slowly to form dew. The glare of a fire on a cold night, or the sun at high altitudes, warms the body of an observer by its radiant energy, though the air itself may be very cold because of its low coefficient of absorption; so when the fire goes out, or the sun goes under a cloud, the air is found to be almost as cold as before.

303. Radiation and temperature. It is a matter of common experience that hotter bodies lose heat more rapidly than cooler ones; also that a cooler body's temperature rises at a more rapid rate than a warmer one when exposed to radiant heat. These effects were examined experimentally by Dulong and Petit, who devised an empirical formula which, however, is valid for only small differences of

temperature. An examination of their observations led Stefan to formulate a law which holds for any range of temperature, and appears to be fundamental for ideal black bodies. It is that *the total rate of radiation emitted by unit area of a black body is proportional to the fourth power of its absolute temperature.*† This may be expressed by the equation $E_1 = kT_1^4$, where k is a constant that has been found to be 5.77×10^{-5} erg per cm^2 per second per $(\text{degree})^4$, when E_1 is measured in ergs per second radiated from one square centimeter. If this body is surrounded by black walls whose absolute temperature is T_2 , lower than T_1 , they also radiate an amount of energy per second given by $E_2 = kT_2^4$. But since the absorptivity of a black body is unity, each surface absorbs all that is radiated from the other, so that the net loss of energy by the hotter one is at the rate of

$$E_1 - E_2 = E = k'(T_1^4 - T_2^4). \quad (1)$$

The constant k' depends upon the areas involved, but if the hotter body has unit area, and T_2 is zero, then the equation reads $E = kT_1^4$, so that k may be defined as the number of calories radiated per second per unit area by a black body whose temperature is one degree above surroundings at the absolute zero.

In the case of bodies that are not perfect radiators, a different constant, k_1 , is called for, which depends upon the nature of the surfaces as well as on their area, and for certain bodies, such as gases, it is not even a constant.

For small temperature differences, an approximate formula may be derived which is often useful. Factoring equation (1) above and taking k_1 as the constant for this case, we obtain

$$E = k_1(T_1 - T_2)(T_1^3 + T_1^2T_2 + T_1T_2^2 + T_2^3).$$

But by hypothesis T_1 is approximately equal to T_2 ; hence, setting $T_1 = T_2$ in the various products of the second parenthesis, we obtain $E = k_1(T_1 - T_2)4T_2^3$, or $E = k_2(T_1 - T_2) = k_2\Delta T$, where k_2 equals $4k_1T_2^3$, assuming T_2 to be constant during the process of cooling. This is known as Newton's law of cooling, which states that the rate is proportional to the temperature difference ΔT , provided this is small.

304. Temperature of the sun. If the sun is regarded as an ideal black body, the temperature of its visible surface, known as the

† A proof of this law due to Boltzmann and based on thermodynamic theory may be found in *A Textbook of Heat* by Saha and Srivastava, p. 519, or in Wood's *Physical Optics*, third edition, p. 799.

photosphere, may be obtained approximately by the use of Stefan's law as follows: If r is the sun's radius and T the absolute surface temperature, then the total heat in ergs radiated per second is given by

$$E = 4\pi r^2 k T^4.$$

At R , the earth's distance from the sun, the same energy is spread over a sphere whose area is $4\pi R^2$, and the amount h falling upon a square centimeter per second at that distance is $E/4\pi R^2$, or

$$h = \frac{4\pi r^2 k T^4}{4\pi R^2}.$$

The solar constant S (Article 294) is measured in calories per minute, while h is in ergs per second; therefore $h = (4.185 \times 10^7 S)/(60)$. The ratio r/R is half of the observed angle which the sun subtends at the earth and equals 4.649×10^{-3} radians. The constant $k = 5.77 \times 10^{-5}$ erg per cm^2 per second per (degree)⁴, as stated in Article 303. Introducing these values, and 1.937 for S , we find the temperature to be 5737° K. This figure is somewhat too small, because the sun does not radiate like an ideal black body and must therefore be hotter than a black body radiating with the same intensity. Actually the sun's surface temperature appears to be about 300° higher than the one just calculated, and in general, "black-body temperatures" of incandescent bodies are lower than their actual values.

305. Radiation pressure. Clerk Maxwell, a celebrated English mathematical physicist, author of the electromagnetic theory of light, showed that as a consequence of the theory, pressure must be exerted on all surfaces exposed to radiant energy. This pressure, acting on an ideal black body, he found equal to the energy of radiation contained in unit volume of the space through which it traveled, provided the exposed surface completely absorbed it, but that it would be twice as great if perfectly reflected. The doubling would also occur if the surface were bombarded with minute corpuscles having inertia, as in the case of an ideal gas whose behavior closely resembles diffuse black-body radiation. If these corpuscles were inelastic and did not rebound, the pressure exerted would be only half the value of that exerted under a perfectly elastic impact.

The actual existence of this pressure was first established experimentally by Lebedew, a Russian physicist, in 1900, and measured quantitatively by both Nichols and Hull in America the following year. Their results agreed with Maxwell's predictions, and greatly strengthened his theory of radiation. The apparatus employed was essen-

tially a radiometer, but contained in such a high vacuum that the effect of bombardment by the gas molecules was reduced to a negligible value. In this case then, the vane having a polished surface experiences a larger pressure than the blackened one, and the resulting torque is opposite to that produced in a poorer vacuum. This torque was measured, and the pressure which caused it calculated.

When light falls upon a very small object, the force it exerts may exceed the force of gravity due to a luminous body like the sun. This is because the total force exerted by the light varies as the section of the particle, or πr^2 , while the gravitational force varies as its volume $\frac{4}{3}\pi r^3$. As the radius decreases, the volume decreases more rapidly than the area, and if the particle is small enough it will be repelled rather than attracted by the sun. This is strikingly illustrated by the tails of comets which are made of minute particles normally surrounding the "head." As the comet nears the sun, these are driven backward to form a tail which streams out in a direction opposite to the sun, and so precedes the comet's head as it again moves away into space.

306. Comparison of radiation and gas pressures. Although they behave differently, there are striking similarities between these pressures. A stream of gas molecules striking an absorbing surface exerts a pressure twice as great as that of a radiant beam of the same energy density. In a cylindrical stream of gas molecules, of unit section, let v be the velocity, m the mass of a molecule, and n their number per unit volume. The time t required to absorb a cubic centimeter is the time required to travel a linear centimeter, or $1/v$. The momentum of the cube is nmv , and the time rate of change of this momentum is nmv/t . Substituting $t = 1/v$, $nmv/t = nmv^2$. This is the force per unit area exerted on the surface, and is equal to the pressure. But the kinetic energy w is $nmv^2/2$ per unit volume, therefore $p = 2w$, whereas in radiation $p = w$. This means that although radiation must have momentum, the momentum is only half as great as that of a stream of particles having the same energy density.

Another striking fact, demonstrated by Professor Hull, is that an enclosure whose walls emit black body radiation is similar to a vessel containing a gas. In both cases adiabatic changes in volume are related to changes in pressure according to $pV^\gamma = b$ (Article 254). The calculated value of γ for a monatomic gas is $5/3$ (Article 214), but with radiation, γ is $4/3$. Therefore, an adiabatic reduction of volume increases radiation pressure, but less so than with a monatomic gas. If a gas is compressed isothermally, its pressure varies inversely as the

volume, but in the case of radiation, there is no change in pressure, because $p = w$, and w is a constant at constant temperature.

367. Radiation and mass. The similarity between radiant energy and mass or inertia is strongly suggested by the preceding discussion. That energy has mass has long been thought probable, because an electric current has inertia by virtue of the energy stored up in the surrounding medium. This consideration led Hasenöhr† to investigate the exact relationship between mass and radiant energy. His theory was based upon an imaginary experiment on an enclosed space filled with ether waves. The energy required to accelerate the container depends upon its inertia, but this inertia is affected by the pressure of the enclosed radiation on its walls. This pressure may be found as explained in the last article. Then by a somewhat complicated calculation, Hasenöhr showed that the mass of energy equals the amount of the energy divided by the square of the velocity of light c , or $m = W/c^2$.

Today the above relation (as shown by Einstein) is an important postulate of relativity, and an essential principle of quantum mechanics. In consequence of this postulate we must regard mass and radiant energy as different aspects of the same thing. In theory they can be interchanged, although we are only beginning to understand how such transformations are brought about. At any rate it is now believed that radiating bodies lose mass, and that extremely hot bodies like the stars, which radiate energy in enormous quantities, must be constantly losing mass.

In the relation $m = W/c^2$, or $W = c^2m$, if we take c as 3×10^{10} cm/sec., and measure m in grams, then W is found in ergs. Thus the energy equivalent of one gram of matter is 9×10^{20} ergs, or 9×10^{13} joules. This is about 34 million horsepower hours, so that the gradual destruction of a single gram of matter would furnish one horsepower for about 3800 years. The sun radiates energy at a rate that may be calculated from the solar constant, or 1.937 calories per square centimeter per minute at the earth. This energy streams across the surface of a sphere whose radius is 93 millions of miles, which means that the sun must be losing mass at the rate of 4.6 million tons per second, but it has so much mass to lose (2×10^{27} tons), that it will be only one per cent less massive after a lapse of over 140 billion years.

† Friedrich Hasenöhr, 1874–1915, an Austrian physicist killed in the World War. His discovery was published in the *Wiener Berichte* in 1904 and in the *Annalen der Physik* in 1904 and 1905.

The preceding calculation applies to all processes in which heat is created or destroyed. It applies, for instance, to chemical reactions. These are usually regarded as subject to the law of the conservation of mass, but actually changes in mass must occur, although they are far too minute to be observed. Thus when oxygen and hydrogen unite to form water, the heat of combination, W , of a gram molecule is 2.86×10^{12} ergs. Its mass equivalent, W/c^2 , is 3.18×10^{-9} g. This is a loss of less than 2 parts in 10 billion of the 18 g produced.

SUPPLEMENTARY READING

- C. F. Eyring, *A Survey Course in Physics* (Chap. 7), Prentice-Hall, 1936.
T. S. C. Northrup, *Science and First Principles* (Chap. 3), Macmillan, 1931.
F. A. Lindemann, *The Physical Significance of the Quantum Theory*, Clarendon Press, Oxford, 1932.
Ingersoll and Zobel, *Heat Conduction with Engineering and Geological Applications*, Ginn, 1913.
T. A. Blair, *Weather Elements* (Chap. 4), Prentice-Hall, 1937.

PROBLEMS

1. How much heat per hour passes through a layer of ice 6 cm thick covering a pool 10 m^2 in area, if the water below it is at 0°C and the air above it at -20°C ? *Ans.* 6×10^6 calories.
2. An aluminum sauce pan has a diameter of 16 cm and its bottom is 4 mm thick. It contains 2 l of water at 20°C and is placed on a stove which maintains its lower surface at a temperature of 160°C . If no heat is lost, how long will it take to boil? How much water will boil away per minute? *Ans.* 6.3 sec.; 1.695 kg. (NOTE: These results do not agree with experience because actually there are very great losses through the sides of the sauce pan and from the upper surface of the water.)
3. An icebox made of oak contains 8 kg of ice at 0°C . Its dimensions are $60 \times 40 \times 40 \text{ cm}$, and the wood is 5 cm thick. How long will it take the ice to melt if the surrounding temperature is 20°C ? *Ans.* 5.79 hours.
4. If the air of a room just inside a glass window $140 \times 80 \text{ cm}$ area and 3 mm thick is at 6°C , and just outside is at 4°C , how many pounds of coal must be burned per hour to supply the loss of heat? (Take 3.5×10^6 calories per pound.) *Ans.* 0.114 lb.
5. If a woolen glove is regarded as being essentially a layer of quiescent air 2 mm thick, and if its total area is 150 cm^2 , how much heat does a person lose per minute from his hand on a winter's day at -5°C ? (Take blood heat at 98°F .) *Ans.* 97.3 calories.
6. Heat flows through a copper bar under a uniform gradient of 2.5°C per cm. If the loss from its lateral surface is negligible, what is the rate of flow per square centimeter of the bar's section? *Ans.* 2.295 calories per sec.

* 7. The rise of temperature, per sec., per cm^3 , of the bar in Problem 6, is found by calculating the thermal capacity of a cm^3 of copper and then dividing the heat available (per sec., per cm^3) by this quantity. Calculate the rate at which any portion of the bar rises in temperature. *Ans.* 2.82°C per sec.

* 8. If the solar constant when the sun is 60° above the horizon is 1.5 calories/ $\text{cm}^2\text{-min.}$ on a horizontal surface, what is its value on a vertical area? How many calories per minute pass through a plate glass window 150×80 cm in area and 2.5 mm thick, if the relative transmission to solar radiation is 12 per cent? *Ans.* 0.866 calories/ $\text{cm}^2\text{-min.}$; 4988 calories per min.

* 9. A jar coated with lampblack is filled with water at 60°C in a room at 20°C . The rate at which it loses heat, by Newton's law of cooling (Article 303) is $k_2\Delta T$, or $40 k_2$. This obviously equals its thermal capacity times the rate at which its temperature falls. If the same jar is now filled with another liquid the rate of *temperature* change is altered, though the rate of *heat* loss is still $40 k_2$. The water cools to 55° in 8 minutes. The other liquid has a density of 0.6 and cools to 55° in 3 minutes. What is its specific heat? *Ans.* 0.625 calories/g.

PART III
WAVE MOTION AND SOUND

CHAPTER 23

Waves

308. Definition. An elastic medium behaves as if it were a succession of adjoining particles between which are balanced forces of attraction and repulsion. If one of these particles is set vibrating, its neighbor follows with the same motion, though lagging behind the first in time. The resulting vibration of the medium is known as **wave motion**, and the instantaneous form of the disturbance as a whole is called a **wave**.

309. A wave-motion model. As an illustration of such a system, imagine a series of metal balls connected by helical springs which resist both extension and compression, thus providing forces of attraction and repulsion between the balls if they are displaced. Then if one ball is moved, its next neighbor moves the same way but a little later, because of its inertia reaction and other forces acting upon it from the yet undisturbed part of the series. The third ball is next displaced from its position of equilibrium, and it in turn acts upon the fourth, and so on. After a time the first returns to its position of rest, followed by the second and the others in succession, and then swings the other way, owing to its inertia, which forces a corresponding motion upon the rest in turn. In the meantime the original disturbance is traveling outward to more distant parts of the system. Thus, while each ball moves only a short distance from its point of rest, the wave travels as far as the system extends, unless its energy is previously dissipated by friction. So the wave advances, but the particles which constitute it do not.

310. Transverse waves. The simplest form of wave motion is one in which the constituent particles perform harmonic vibrations at right angles to the line connecting them, which is the direction of the wave motion.

We have already seen that the various positions of such a vibrating particle, if plotted with time as the axis of abscissas, form a sine curve resembling a wave. Now if the successive particles in a long straight row perform the same vibrations, but each lags a definite time behind its neighbor, then the points of the sine curve represent correctly the position of the particles at a given instant of time. Such a curve is

then an instantaneous picture of a wave traveling at a constant speed, and the X axis measures distance rather than time, as in Article 92. This is because the phases of the successive particles at a given instant

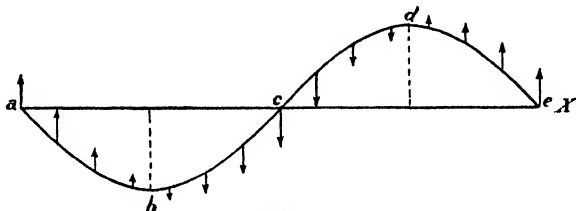


Fig. 1.

vary with the *distance* from the origin, while in Article 92, only one particle is considered, and its phase depends only upon *time*. Suppose the wave, shown in Fig. 1, originates in the harmonic vibration of the particle a which has just completed a cycle and is starting upward again with the maximum velocity which belongs to its mid-position. The front of the wave has reached the particle e , which is just starting upward, and the intermediate particles have instantaneous velocities as indicated by the varying length of the arrows. The particle c has completed half a cycle and is just starting downward, while b and d are at the lower and upper ends of their paths, having zero velocity but maximum acceleration.

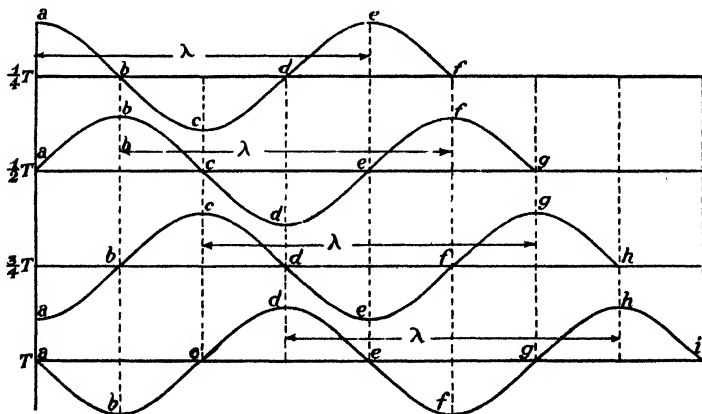


Fig. 2.

If a continues to vibrate, it sends out a succession of such waves, and their subsequent appearance at the end of each of the next four quarter periods is shown in Fig. 2. A study of these diagrams shows

how the wave front advances, picking up new particles, and causing them to vibrate in turn, while those behind it are in various phases of their cycles. As we trace the wave *backward* toward the origin, each particle is in a *later* phase than the one to the right of it, so that *a* represents the latest or *youngest* aspect of the wave, just starting on its career, and *i*, in the lower curve, is in the phase which originally started from the origin two periods earlier. It is therefore the *oldest* portion of the wave, just as persons born earliest in a community are the oldest.

For convenience, regions like *cde* and *ghi* in the lowest curve are called *crests*, while *abc* and *efg* are called *troughs*.

311. Wave length and velocity. *The length of a wave is the distance measured parallel to the line of propagation between the two nearest particles that are in the same phase at the same time.* By *phase* is meant not only position, but direction of motion, of a vibrating particle. Thus in any of the four curves of Fig. 2 the particles *a* and *e* are always in the same phase, doing the same thing at the same time. So are *b* and *f*, *c* and *g*, and so on. But *b* and *d* are not in the same phase because one is going up while the other is going down. So *c* and *e* as well as *d* and *f* are said to be in *opposite phase*. Therefore the horizontal distance between *a* and *e* is a wave length, as well as the distance between *b* and *f*, *c* and *g*, and so on, as indicated in the diagram.

The velocity with which a wave travels depends upon the nature of the medium, its elasticity, density, and so forth. It may be calculated for some simple cases, as we shall see later, but the *velocity is always equal to the wave length times the frequency*, no matter what the medium or the mode of vibration may be. This is written $v = \lambda n$ where λ is the wave length and n the frequency, a relation which is really self-evident. Imagine watching waves of the ocean rolling inward along a pier. If we can estimate their length λ by comparison with a measured distance on the pier, and count the number N that pass in a minute, then their velocity is obviously λN feet per minute, or $v = \lambda N/60 = \lambda n$ feet per second. The Greek letter *lambda* is always used to express wave length, a custom that offers the only difficulty in a very elementary relation which we recognize intuitively. This expression may also be written $v = \lambda/T$, where the period T is the reciprocal of the frequency. In this form it is just as obvious and sometimes more convenient.

312. Examples of transverse waves. The most important example of transverse waves is the propagation of light and radiant heat,

although the mechanism by which such waves are carried through empty space is still far from clear. However, when they are traversing transparent matter such as the lenses and prisms of optical instruments, their transverse wave character is well established. The much longer waves used in radio transmission are of the same sort, and in this case the mechanism of their production, by oscillatory electrical currents, furnishes additional evidence, not only that they are transverse waves, but also of their electromagnetic character.

Transverse waves can be set up mechanically in ponderable matter, provided it possesses elasticity of form, like solids. Gases cannot sustain transverse vibrations because they have no form rigidity, and liquids can do so only on their free surface, in a manner to be discussed later. A long stretched wire has mass and elasticity, and transverse waves can be produced in it by setting some portion of it in vibration. They travel in general too fast for the eye to follow, but a long rubber tube filled with sand, when only slightly stretched, carries the wave slowly enough to allow it to be seen. A quick up-and-down motion of the hand holding one end sends a crest traveling along the tube toward the other end.

313. Longitudinal waves. If a succession of particles in an elastic medium vibrate in the direction of the line joining them, instead of transversely, a longitudinal wave is set up in which each succeeding particle lags slightly in phase behind its predecessor, as has already been explained. This may be illustrated by a row of metal spheres separated by helical springs, such as was described in Article 309 as a model of waves in general. These are supposed to be supported by long threads whose influence on their motion is negligible. The row numbered 1 in Fig. 3 shows the system at rest. The spheres are subjected to attraction when the springs are stretched and to repulsion when they are compressed, as well as to inertia because of their mass. In all these respects they experience influences exactly like those that act upon the molecules of mediums having volume elasticity, such as gases, liquids, and elastic solids, although the mechanism by which this effect is achieved is of course not the same.

Let the particle *a* be displaced from its position of equilibrium, shown in row 1, to the position occupied in 2. This compresses the spring between it and *b*, driving *b* also to the right, but not immediately so far, because of its inertia and because its motion is opposed by the spring between it and *c*. Now *a* is acted on by the pull of the spring behind it (assuming the row continues to the left), as well as a push from in front, and it returns during the next quarter period

to mid-position as shown in row 3, while the inertia of b has carried it much nearer c , displacing the latter in turn. In row 4, a quarter period later, the inertia of a has carried it backward to a displacement nearly as great as in 2, while b , under the action of the springs, is in mid-position with a reversed momentum which carries it backward to a new position shown in row 5.

A study of this series of diagrams shows that the advancing wave is marked by a compression, or *condensation*, between the two first

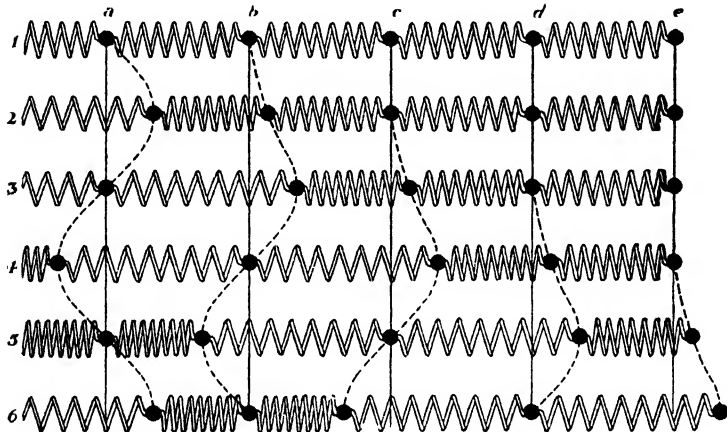


Fig. 3.

particles indicated by the compressed spring in row 2, which is then followed by a *rarefaction* indicated by the extended spring. This again is followed by a second condensation which appears at first in row 5 and which has advanced one space in row 6. Thus such a wave advances, not by a succession of crests and troughs, but by a succession of condensations and rarefactions.

Of course, such a machine as the one described above is a very crude illustration of a medium whose particles are as nearly continuous as those of most ponderable bodies, including even gases under ordinary pressures. Actually in the transmission of such a wave through matter, the compressions shade off gradually into the rarefactions, so that it is not easy to say just where each begins and ends.

314. Graphic representation. Since the forces of restitution acting upon the particles of a perfectly elastic medium are proportional to their displacement, their vibrations are harmonic and may be represented by sine curves. In a succession of such curves, each in slightly later phase than the preceding one, and placed side by side,

we may obtain as accurate a picture of a longitudinal wave as we wish, depending only upon the closeness of the particles whose vibration is represented. In Fig. 4 the succession of different particles is

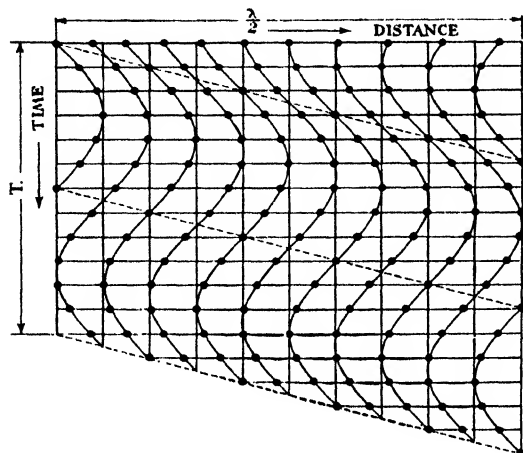


Fig. 4.

shown by dots along any one of the horizontal lines, at any instant. The sine curves drawn with the vertical lines as their time axes show the succession of positions of a single particle during successive time intervals. The progress of a compression followed by a rarefaction is seen as we examine the successive horizontal rows, each row constituting, as it were, one of the instantaneous photographs of a cinema film.

The time that elapses between one compression (or condensation) and the next, at the same place, is evidently a period, while the distance between two adjacent compressions (or condensations) is equal to the wave length, only half of which is shown. We may also see from a study of the diagram that the particles comprising a condensation are all moving in the direction in which the wave is traveling (left to right), while those in a rarefaction are moving in the opposite direction.

It is so difficult to exhibit a longitudinal wave by a graph, that it is usually represented as a transverse wave, or sine curve. When this

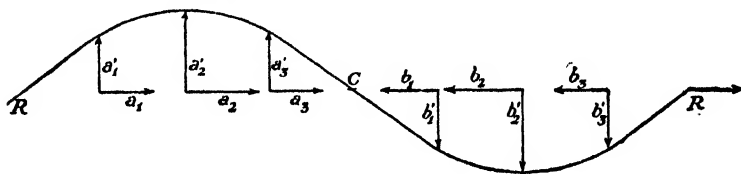


Fig. 5.

is done, it is always understood that the positive displacements a' (Fig. 5) are really displacements a in the direction of wave motion, while negative displacements b' are really displacements toward the

source of the wave. This means that every alternate point of zero displacement is one of maximum condensation C at that instant, while the other points R of zero displacement are those of maximum rarefaction. Thus a longitudinal wave has the same wave length as its equivalent sine wave; that is, from R to R , or C to C . In either case it is the distance between the two nearest particles that are doing the same thing at the same time.

315. Equation of wave motion. From the diagram of Fig. 4 it is evident that the displacement of any of the particles that constitute the wave, whether it is longitudinal or transverse, depends upon both the time and the distance from the origin. The displacement in harmonic motion as a function of the time has been shown (Article 91) to be given by

$$y = r \sin \omega t = r \sin 2\pi t/T. \quad (1)$$

But this refers to only a single point where the displacement is zero when time is zero, or at the beginning of any period. At some other point distant x from this origin of the waves, the phase of a vibrating particle is always *earlier* than at the origin where the *latest* phase is starting. To find this displacement at x , we must deduct from t , in equation (1), the time the wave takes to travel the distance x . This time is x/v , so that equation (1) becomes

$$\begin{aligned} y_{x,t} &= r \sin 2\pi \left(\frac{t - x/v}{T} \right) \\ &= r \sin 2\pi \left(\frac{t}{T} - \frac{x}{vT} \right). \end{aligned}$$

But $vT = \lambda$; therefore

$$y_{x,t} = r \sin 2\pi \left(\frac{t}{T} - \frac{x}{\lambda} \right). \quad (2)$$

This is the equation of an harmonic wave (either longitudinal or transverse) traveling in the positive direction, where the phase angle θ is $2\pi[(t/T) - (x/\lambda)]$ at a point whose distance from the origin is x , and when the time from the beginning is t .

316. Gravitational waves on liquid surfaces. There are two causes for waves on liquid surfaces. One is gravity, and the other is surface tension. In gravitational waves the weight of the liquid supplies the force of restitution necessary to produce a vibration, though in this case it is not simple harmonic. If a depression is made in the free surface of a liquid, as by lowering an object into a pail of water and then withdrawing it, the depressed layer tends to regain its normal level. This is due to the unbalanced pressure acting upon it from

all sides created by the temporary deficit, as shown in Fig. 6 (a). If the object *A* is withdrawn rapidly, the inertia of the entire mass of liquid causes the central portion to overshoot the normal level *L*, while the outer portion sinks below it, as shown in Fig. 6 (b). This

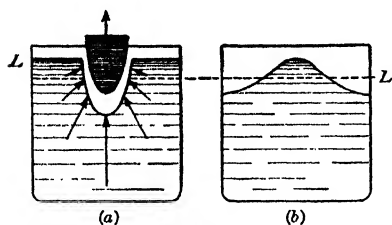


Fig. 6.

results in an up-and-down oscillation which may be regarded as the source of gravitational waves.

If instead of a pail, a long trough had been used, this disturbance would travel along it, while if created at the center of a pond, circular waves would travel out in all directions. Such waves, when the

water is sufficiently deep, can be shown to be the result of two equal harmonic vibrations at right angles to each other. These produce circular vibrations of the particles moving in vertical planes. They are thus simultaneously performing transverse and longitudinal vibrations, but with a ninety-degree difference of phase. The particles in their circular orbits are each in an earlier phase of motion as we advance in the direction of propagation, as is seen in Fig. 7. The curve drawn through these points is not a sine curve, because it is evident that the instantaneous positions of the particles are closer together (horizontally) in the crest of the wave than in the trough, owing to the "condensation" caused by the longitudinal component of the vibration. This makes such waves resemble a succession of steep mountains with broad valleys between, and the liquid moves *with* the wave on the crest, but *against* it in the trough.

In shallow water, whose depth is less than the wave length, the orbits of the particles become elliptical, with their major axes horizontal and their vertical axes diminishing with their depth, until at the bottom the motion is purely horizontal, as we should expect.



Fig. 7.

The tendency of deep sea waves to "break" is due to the fact that the water in the crest is moving forward, while that in front of it, in the trough, is moving backward. This constitutes a sort of shear tending to make the upper layers slide over the lower ones. The wave actually begins to break when the water of the crest is moving

as fast as the wave itself. This occurs if the amplitude is sufficiently large compared to the wave length, and the resulting condition is shown in Fig. 8, where the maximum horizontal component of the orbital velocity is v_o , and the wave velocity is v_w . If h is the height of the crests above mean sea level, it is also the radius of the orbits of the particles; therefore

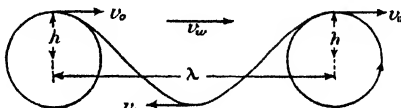


Fig. 8.

$$v_o = \frac{2\pi h}{T} = 2\pi h n, \text{ and } v_w = \lambda n.$$

Hence the condition for breaking is

$$2\pi h n = \lambda n, \text{ or } h = \frac{\lambda}{2\pi}. \quad (1)$$

Crests then do not break (unless a strong wind forces them to) until their height is nearly one sixth of the distance between them.

The velocity of gravitational waves is independent of the density, for the same reason that a pendulum's period is independent of its mass, and its value can be proved equal to

$$v_g = \sqrt{\frac{g\lambda}{2\pi}}, \quad (2)$$

where λ is the wave length and depends upon the frequency according to the usual relation $v_g = \lambda n$. Therefore, substituting for v_g above, we find that

$$\lambda = \frac{g}{2\pi n^2}. \quad (3)$$

317. Surface-tension waves. We have seen that the pressure under a curved surface film increases with its curvature. This pressure under the surface of short waves, called ripples, acts in a manner similar to gravity on long waves, but it is due instead to the elastic behavior of the surface membrane, and tends to smooth out the surface with an intensity which, for very short waves, far exceeds the pull of gravity.

Since the surface tension supplies the force of restitution, as gravity does in the case of long waves, we should expect the velocity of short waves to depend upon the value of that tension, just as v_g depends upon g . This is actually the case, and it can be proved that

$$v_T = \sqrt{\frac{2\pi T}{\lambda d}},$$

where T is the surface tension in dynes per centimeter and d is the density of the liquid. The fact that λ is in the denominator shows that short ripples travel faster than long ones, in contrast to long waves whose velocity increases with the square root of the wave length, as is evident from equation (2) in Article 316.

318. Comparison of waves and ripples. The distinction between waves and ripples is not a sharp one, and in the border land, with λ between 10 cm and 3 mm, the effects of both surface tension and gravity have to be considered. A comparison of the two expressions for velocity shows that there must be a certain critical wave length which moves more slowly than any other over the surface of a given liquid, because v_g decreases with shorter waves, while v_T decreases with longer ones. This minimum occurs when the two velocities are equal, or

$$\frac{g\lambda}{2\pi} = \frac{2\pi T}{\lambda d},$$

whence

$$\lambda = 2\pi\sqrt{\frac{T}{gd}}$$

In the case of pure water at 15°C , $T = 73$ dynes/cm approximately, and $d = 1$; therefore $\lambda = 2\pi\sqrt{73/980} = 2\pi \times 0.273 = 1.715$ cm. This corresponds to a velocity of 23 cm/sec., which is obtained from the equation combining both effects, when the resultant velocity is given by $v^2 = v_g^2 + v_T^2$.

319. General properties of waves. Waves of all types may be reflected, refracted, diffracted, and made to interfere with each other. **Reflection** occurs at the bounding surface between two media which differ in one or more physical properties, such as density, elasticity, and so on. The reflected wave is formed at the interface, or bounding surface, between the two media, and travels back into that which carried the incident wave. **Refraction** is an effect due to a change in the velocity of propagation of the wave when it passes through the bounding surface between two media. **Interference** occurs when two waves either reinforce or reduce each other according to whether the two vibrations they set up at a given point are in the same or opposite phase. It is closely associated with **diffraction**, which is the bending of waves around obstacles, and will be more particularly discussed in the section on light.

320. Huygens' principle. Though devised to help explain a wave theory of light, this principle is applicable to all kinds of waves, and

is extremely useful in simplifying the theory of many wave phenomena. It was published by the Dutch philosopher Huygens of Leyden in 1690 in a treatise on light considered as a wave motion, but was not then generally accepted because of the great prestige of Newton, who advocated a corpuscular theory. This was not finally replaced by the wave theory until 1818. When a vibrating source O (Fig. 9) sends out waves, the locus of the points $p_1, p_2, p_3, \dots p_n$, which the disturbance reaches in the time t , is called a wave front. In a homogeneous region around a vibrating point, or sphere, this wave front is a sphere, but it is a circle if the motion spreads out in only two dimensions, as over the surface of a liquid.

Huygens affirmed that each of the points in the wave front might itself be regarded as a source of waves propagated *beyond their common locus*, and that at any later time the position of the new wave front was the envelope of the wavelets sent out by all of the particles in the old wave front. Thus in Fig. 9 the circles E_1, E_2 , and E_3 are the successive envelopes of the smaller circles that originate in the points p at the end of equal intervals of time.

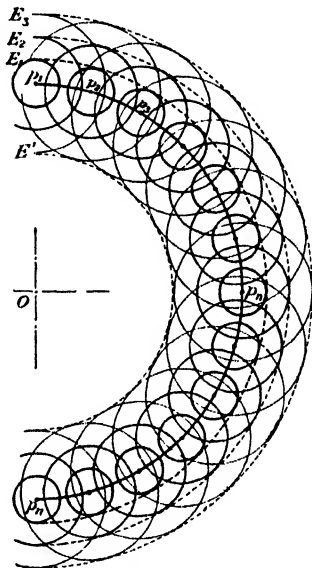
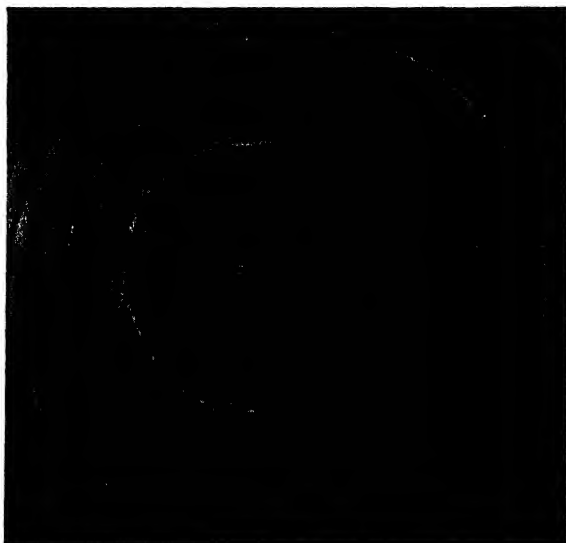


Fig. 9.

There are two difficulties with this principle. One is to account for the condition of the medium between the original and the new wave front. This was explained by Stokes, who proved that the elementary wavelets destroy each other by interference, except along their common envelope. The other difficulty is in explaining why the points p do not send a wave backward toward the origin, which would be the case if they were genuine sources. Then the wavelets would form complete circles (or spheres) as indicated by the inner portions of the small circles, whose envelope is E' , traveling toward O . Of course this does not occur unless the points are the actual source of vibration. The wave *does* travel outward and not backward, and Huygens' principle takes this fact into account, and applies only to the medium lying beyond an instantaneous wave front in the direction of propagation.



Courtesy Professor A. L. Foley, Indiana University.

Plate 1.

Photograph of sound wave from a point source illustrating Huygens' principle. Wavelets both transmitted and reflected by circular "grating" are shown.

321. Reflection. The various wave phenomena are most easily explained by Huygens' principle. In the case of reflection, let us consider the special case of a source a long way off. Then the wave fronts are practically plane surfaces which appear as straight lines in the diagram and are called plane waves. Thus in Fig. 10 the incident

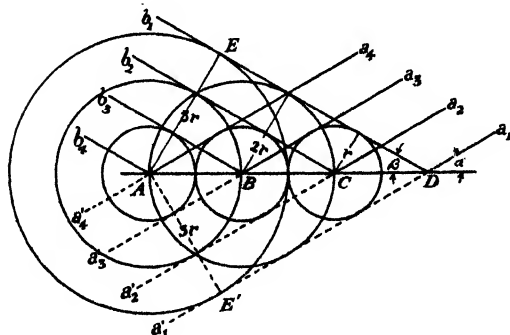


Fig. 10.

wave fronts of a series of waves are shown as a_1 , a_2 , a_3 , and a_4 , separated by a distance r . These would continue to a'_1 , a'_2 , a'_3 , and a'_4 , if there were no interface. But when a_1 reached A , that point became the source of spherical waves which appear circular in the dia-

gram. As a_1 has advanced a distance $3r$ since it first touched A , and is now at D , the disturbance from A which travels with the same speed has

developed a spherical wave front whose radius is $3r$. Then b_1 , the tangent to this sphere at E and passing through D , is the reflected wave front. This is also tangent to a sphere, centered at B with a radius $2r$, which started when a_1 touched B , so it is the envelope of all the wavelets sent out from points in the interface.

The smaller spheres not mentioned above may be regarded as caused by the other wave fronts a_2 , a_3 , and a_4 , in an exactly similar manner,



Courtesy, Professor A. L. Foley, Indiana University.

Plate 2.

Photograph of sound wave from point source reflected from a plane surface.

giving rise to the reflected wave fronts b_2 , b_3 , and b_4 . The angle α of the incident waves is known as the **angle of incidence**, and β is the **angle of reflection**. These are equal because the right-angled triangles ADE and ADE' have one side in common, and $AE = AE' = 3r$ by construction. Therefore the corresponding angles ADE and ADE' are equal, and $ADE' = \alpha$, being vertical angles.

322. Rays and images. A ray in an isotropic medium is any line drawn perpendicular to the wave front in the direction of propagation. In spherical or circular waves it starts at the origin and is a radius of the wave surface. Images due to a plane surface may be

located by making use of rays, and of the law that the angle of incidence equals the angle of reflection. In Fig. 11 the ray Oa from the source O is perpendicular to a wave front at the point p . The tangent at p makes an angle α with AB , and this equals the angle Oab

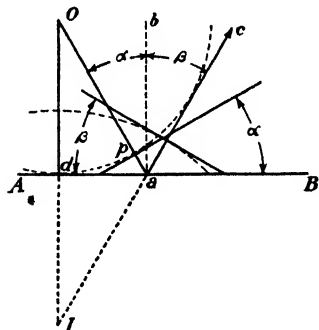


Fig. 11.

between the ray, and ba drawn perpendicular to the surface, because their sides are mutually perpendicular.

The wave reflected at a is defined by the ray ac , which makes the angle cab with ab , and this angle is obviously equal to $\beta (= \alpha)$, between AB and the tangent to the reflected wave front normal to ac . The origin of the reflected waves must lie on the line ac or ac produced. Similarly the ray Od is reflected back upon itself, since the angle of incidence, and therefore of reflection, is zero

degrees. This fact locates the origin as lying on Od produced, so that the image of O must be at the intersection of Od and ac , or at I . But $\angle dOa = \angle Oab$, since they are opposite interior angles, and $\angle dIa = \angle bac$ (sides parallel). Then the right-angled triangles Oda and $I da$ are equal, having one side da in common and the angles opposite that side equal. Therefore $Od = Id$, which means that *the source of the waves, and its image are equidistant from the interface, which is at right angles to the straight line joining them.*

323. Change of phase at the interface. Longitudinal waves. When a longitudinal wave is reflected from a surface or a medium that tends to impede the vibrations of the particles more than the medium in which it was moving, a condensation is reflected as a condensation and a rarefaction as a rarefaction. This is easy to account for with the aid of the model already made use of in Article 309, but with the system of balls and springs attached to a solid wall. In Fig.

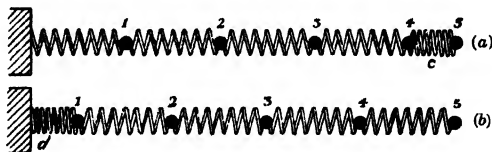


Fig. 12.

12 (a), a compression has been started by forcing the two outer balls together. This is passed rapidly along the line to where the spring is compressed against the wall, which does not move, as shown in (b). This compression is therefore reflected, forcing 1 out

again in phase opposite to its original motion. This particle acts upon 2, and so on down the line, a compressional wave having resulted from one of the same kind. Similarly, a rarefaction created by pulling 5 away from 4 would ultimately result in pulling 1 away from the wall, but the wall cannot follow, and its reaction on 1 results in a recoil in phase opposite to its original motion. This pulls it away from 2, causing a rarefaction that is passed back again to the farther end.

If the reflection is from a surface beyond which vibrations take place more freely than in the first medium, a condensation is reflected as a rarefaction, and vice versa. This may be seen with the same model, if we suppose the system laid out on a frictionless surface with both ends free. A compressional wave traveling down the line from left to right causes 5 to fly off farther than the other balls. This reflection then involves motion without change of phase. There is a rarefaction between 5 and 4, followed by one between 4 and 3, and so on back to the origin of the disturbance. On the other hand, a rarefaction started by pulling 1 to the left results in a pull on 5, which responds more readily than any of the others because no spring exerts a tension on it from the right. It therefore compresses the spring between it and 4, as its inertia carries it without change of phase beyond its normal distance from that ball, and this compression travels back to 1 as the reflection of the original rarefaction.

324. Change of phase at the interface. Transverse waves. In Fig. 13 we see the two cases of the reflection of transverse displacements. A cord hanging from a rigid support is shown in *a* as having had a hump impressed upon it by a quick lateral jerk of the free end. This hump has almost reached the support and is about to be reflected. A sidewise force acting to the left now tends to move the support that way, but as it resists, the cord recoils under the reaction, the displacement reverses phase, and comes down the cord, as shown in *b*.

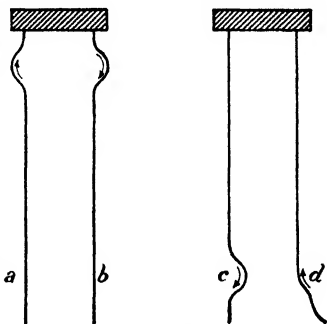


Fig. 13.

In *c*, a similar hump started at the top has just reached the bottom, and there, because of the inertia of each succeeding portion of the cord, it encounters no resistance such as it met higher up, so it flies out farther than ever, as seen in *d*, and starts a reflected wave in the same phase back toward the support.

We may summarize the foregoing conclusions as follows, using the term *dense* for properties that impede vibration and *rare* for properties that help it:

When longitudinal waves are reflected against dense back into rare, there is a reversal of the phase of the particles, and condensations and rarefactions are unchanged, but in reflection against rare back into dense, there is no change of phase of the vibrating particles, but condensations are reflected as rarefactions, and vice versa.

When transverse waves are reflected against dense back into rare, there is a reversal of phase, a crest is reflected as a trough, and vice versa, but in reflection against rare back into dense, there is no change of phase, and both crests and troughs are reflected unchanged.

Long gravitational (water) waves, strictly speaking, are both transverse and longitudinal, and the conditions of reflection are unlike a pure transverse wave. On meeting an obstacle like a sea wall, a crest is reflected as a crest, and a trough as a trough.

325. Interference. When two waves meet or cross each other at an angle, those portions which are in opposite phase at any point tend to neutralize each other and are said to *interfere*. This can be accomplished with water waves, sound and light, or any electromagnetic wave. Sound and light will be discussed as acoustic and optical phenomena, but we shall now examine the interference of water waves, which is typical of all other kinds.

If an electrically driven tuning fork has wires fastened to both prongs at right angles so that they dip below the surface of a sheet of water under the horizontal fork, the vibrations set up in this way radiate from both points in concentric circles which cross each other and produce a striking interference pattern, shown in Fig. 14. The two vibrating wires, *A* and *B*, are the centers of concentric waves whose crests are indicated by solid lines and the troughs by dots. The vertical line *a*, bisecting at right angles the line joining *A* and *B*, is a locus of reinforcement where crest meets crest and trough meets trough. This is because every point in this line is equidistant from the two sources, and the portions of the wave fronts which approach it are sent out simultaneously in the same phase by the prongs as they alternately approach and recede from each other. The next two lines, *bb*, are continually loci of interference as indicated by the intersection of the crests and troughs through which they are drawn.

Any point on these curves is farther from one origin than the other by half a wave length, so that at any instant the disturbance from *A* must be in opposite phase from that due to *B*. If we call these distances d_A and d_B , then the lines *bb* are defined by the equation $d_A - d_B = \lambda/2$. Since this is the well-known definition of an hyperbola, the curves are hyperbolic, and *A* and *B* are the foci.

In like manner the curves *cc* are hyperbolic loci of reinforcement defined by $d_A - d_B = \lambda$. The curves *dd* are loci of interference defined by $d_A - d_B = 3\lambda/2$. So in general the curves denoting reinforcement are defined by $d_A - d_B = 2n\lambda/2$, where *n* is any integer, so that $2n$ is always even, and $2n\lambda/2$ represents a succession of path differences equal to multiples of whole wave lengths $\lambda, 2\lambda, 3\lambda, \dots n\lambda$. Similarly, curves denoting interference are defined by $(2n - 1)\lambda/2$, where $2n - 1$ is always odd, and the whole expression gives the successive path differences equal to $\lambda/2, 3\lambda/2, 5\lambda/2, \dots (2n - 1)\lambda/2$, or an odd number of half wave lengths.

The interference of two waves does not involve loss of energy, as might be expected. The fact that the amplitude is much reduced, if not destroyed, wherever the waves meet in opposite phase, is counterbalanced by an increased amplitude in regions of reinforcement. If the two waves have the same amplitude, the resultant amplitude is zero where they interfere, but twice that of either where they meet in the same phase. This actually means four times the energy of either wave acting independently at such points, while the energy of the system as a whole may be proved to be equal to the sum of the energies of the two waves taken separately.

The interference of two waves does not involve loss of energy, as might be expected. The fact that the amplitude is much reduced, if not destroyed, wherever the waves meet in opposite phase, is counterbalanced by an increased amplitude in regions of reinforcement. If the two waves have the same amplitude, the resultant amplitude is zero where they interfere, but twice that of either where they meet in the same phase. This actually means four times the energy of either wave acting independently at such points, while the energy of the system as a whole may be proved to be equal to the sum of the energies of the two waves taken separately.

326. Stationary waves. The lines *a, b, c*, and so on, defined above, intersect the line joining *A* and *B* at a row of points as indicated in Fig. 15, and we may confine our attention to what goes on at these points without considering the hyperbolae as a whole. Starting at *A*, which is always in vibration, we come to the point *d* on a curve denoting interference where the disturbance is a minimum. At *c* the

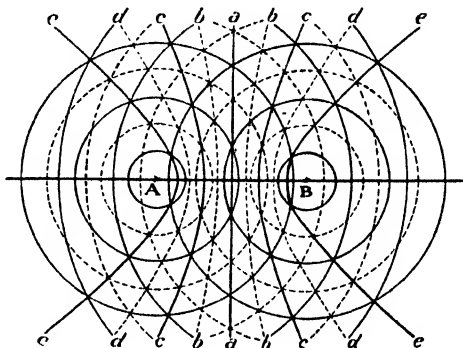


Fig. 14.

vibration is once more a maximum, at b again a minimum, and at a another maximum. Moreover if two troughs meet at a , two crests must be meeting at c , and so on, so that a is always in opposite phase to either point c . A view of this wave formation taken from E looking

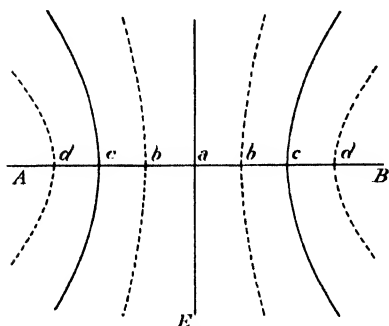


Fig. 15.

along the surface would appear like Fig. 16 (a). If the situation were as just supposed, but half a period later, Fig. 16 (b) represents the case, with a crest at a instead of a trough, and with troughs at c . These are called **stationary or standing waves**, because they do not advance, but vibrate up and down with alternate segments in opposite phase. Between these maxima are points like b and d , which vibrate very little or not at all. These are called **nodal points**, or **nodes**, while the segments between them, like bd , are called **loops**, and their centers a , c , and so forth, **antinodal points**. These terms are applicable only to standing waves, and should not be confounded with the crests and troughs of traveling waves. A loop may be either a crest or a trough at a given instant, while a node does not vary, remaining fixed both in time and space. The term should therefore not be used to describe the point midway between crest and trough of a traveling wave, because this point moves along with the wave, and is in no sense at rest.

Longitudinal waves of the same frequency, and traveling in opposite directions, also produce standing waves. As has been noticed in Article 314, the particles of a condensation are traveling *with* the wave, and those of a rarefaction, *against* it. Therefore when condensation meets condensation, the motions of the particles, being opposed, neutralize each other, and so for rarefaction meeting rarefaction. But when a rarefaction meets a condensation traveling in the opposite direction, the motions of the particles are in the same sense. They reinforce each other, and the result is an increased amplitude of vibration. Thus a node is the point at which condensations meet

ing along the surface would appear like Fig. 16 (a). If the situation were as just supposed, but half a period later, Fig. 16 (b) represents the case, with a crest at a instead of a trough, and with troughs at c . These are called stationary or standing waves, because they do not advance, but vibrate up and down with alternate segments in opposite phase. Between these maxima are points like b and d ,

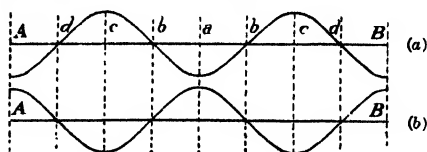


Fig. 16.

condensations, or rarefactions meet rarefactions, while an antinode is the point at which condensations meet rarefactions.

327. Stationary waves caused by reflection. In the case of gravitational water waves meeting a solid wall, the reflection is without change of phase, while in a stretched string the transverse vibrations are reflected in opposite phase. In the first case the reflecting surface is a region of maximum vibration, while in the second it is a region of zero vibration, or a node. But in both cases we have waves of the same length traveling with the same velocity in opposite directions, which is the condition necessary for creating stationary waves. Similarly, longitudinal waves may be reflected and so develop standing waves through interference between the original and reflected waves. If they are reflected against a denser back into a rarer medium, the reversal of phase at the interface results in destructive interference and there is a nodal region. Half a wave length away from the interface the original and reflected waves are in the same phase and there is an antinode.

To sum up, longitudinal and transverse standing waves, caused by reflection against dense back into rare, have nodes at the interface. If against rare into dense, they have antinodes at the interface. If water waves are reflected against a solid obstacle, the interface marks an antinode. This is easily seen when ocean rollers are reflected from a sea wall. The "choppy water" which results from the interference of the two wave systems constitutes a standing wave and has its greatest vertical motion at the wall.

328. Refraction. When a train of waves reaches an interface between two media in each of which it travels with a different speed, the wave front is swung around and the direction in which it is moving is altered. This is strikingly represented by the waves of the ocean, which, when well off shore, may be inclined to it, as shown in Fig. 17, where α is the angle between the wave front and the line of the beach AB . But as the waves get into shallow water at ab , they travel more slowly and ultimately approach the beach in lines almost parallel to it. This is because the part of the wave indicated by p reaches shallow water first, and is retarded before q , which keeps on at its old speed for a time. This tends to swing the wave front around to face the beach, and its whole direction is altered.

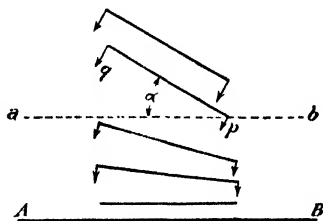


Fig. 17.

The laws of refraction may be readily derived from Huygens' construction as follows: The plane wave front ABC is shown in Fig. 18 just reaching the interface between the two media Q and R . In Q it travels a distance $BB' = CC'$ per second, while in R it moves more slowly over a distance r per second, where r is less than CC' . During

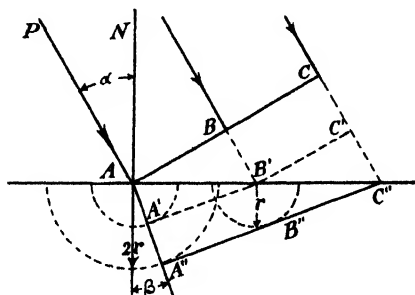


Fig. 18.

the two seconds it takes for C to reach C'' , the disturbance, regarded as starting at A , travels a distance $2r$, the radius of the spherical wave drawn about A as a center. Similarly, after the lapse of one second, B' becomes the center of another spherical wave of radius r . Now according to Huygens' principle the envelope of these wavelets,

as well as others which proceed in the same manner from other points between A and C'' , is the new wave front. It is therefore a plane represented by the line $A''B''C''$ drawn tangent to the circles, and is said to have been refracted.

The angle α between AC and the interface is the angle of incidence, and β between $A''C''$ and the interface is the angle of refraction. Or these angles may be defined with reference to a line N drawn normal to the surface, so that α and β are the angles between this normal and the rays PA and AA'' of the incident and refracted wave systems.

329. Law of refraction. Now consider the triangles ACC'' and $AA''C''$ in Fig. 18. Both are right angled, one at C and the other at A'' , by construction. Therefore $CC'' = AC'' \sin \alpha$, and $AA'' = AC'' \sin \beta$. Dividing the first of these equations by the second, we obtain $CC''/AA'' = \sin \alpha/\sin \beta$. But CC'' is equal to the distance traveled by the original waves in the time t (assumed as two seconds) with the velocity v_1 , and AA'' is the distance traveled in the same time by the refracted system with a lower velocity v_2 . Therefore

$$\frac{CC''}{AA''} = \frac{v_1}{v_2} = \frac{\sin \alpha}{\sin \beta},$$

or the velocities before and after refraction are to each other as the sines of the angles of incidence and refraction respectively. This ratio of the velocities in the two media is of the greatest importance in the theory of light. It is known as the index of refraction, and is

usually designated by the letter n . It is greater than unity when $v_1 > v_2$, as assumed above.

If the second medium R had permitted a higher velocity than the first, then r would be greater than CC' , the wave would have been swung the other way, and a given ray, instead of bending toward the perpendicular, as shown above, would have been bent away from it. In this case the index n would be less than unity, because v_1 would be less than v_2 . However it is always customary to measure the index on the assumption that the waves are entering a denser medium, which really defines n as the ratio of the higher to the lower velocity in all cases. The index of refraction is often defined as the ratio of the sine of the angle of incidence to the sine of the angle of refraction, but this should be qualified by specifying the direction as from "rare" to "dense."

330. Free vibrations and resonance. The vibrating source of the waves we have been considering has in general a natural frequency like that of a pendulum or tuning fork, and when allowed to vibrate in its own way, it is said to perform **free vibrations**. In order to maintain these vibrations and the waves which they send out, a periodic force must be applied as in a clock, where the escapement causes the spring or weights to give an impulse during every oscillation. If this is timed correctly, only a very small force is necessary to maintain the vibration, since it has to supply only the energy lost in friction during each period. Thus a small impulse applied during the first part of each swing is sufficient to keep a pendulum or balance wheel executing free vibrations indefinitely. If the impulse were increased even a little, the amplitude would build up apparently out of all proportion to the cause, because the energy in excess of that required to take care of friction losses is cumulative, and is continually stored up, as it were, in the increasing energy of vibration. In this way prolonged repetition of a small periodic impulse may produce astonishing results due to what are called "superimposed vibrations." The rhythmic marching of troops over a suspension bridge sets it swinging if the natural period of the bridge is an integral multiple of that of the marching tempo. Then the impulses always occur in the same phases of the swing of the bridge, and it oscillates more and more violently till its amplitude assumes dangerous proportions. On this account the order to break step is usually given to troops when marching over such a bridge. In the same way portions of a building may be set in strong vibration by the continued production of one or more particular tones of an organ.

Wherever a series of periodic impulses coincide in frequency with the frequency of a vibrating body, the two are said to be in **resonance**. This word is used particularly to apply to sound when one sounding body causes another having the same natural period as the first (that is, in resonance with it) to develop what may be called **sympathetic vibrations**.

331. Damped vibrations. If a pendulum or other vibrating body is left to itself after having been set swinging, its vibrations tend to die

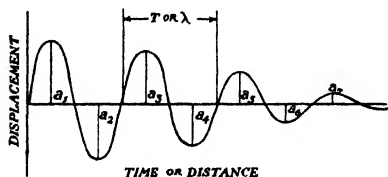


Fig. 19.

down as a result of friction, air resistance, and so on. This gradual decrease of the amplitude results in what is known as a **damped vibration**, and if such a body is sending out a train of waves, they are damped also, and their amplitude gradually grows smaller with

the time. A linear train of waves, caused by an undamped vibrating source, may be damped themselves as they advance. This occurs when the medium is not perfectly elastic and so uses up energy. The amplitude of the waves then diminishes progressively, as signals in an ocean cable, which grow fainter at increasing distances from the source. Damped vibrations or a damped wave train may be represented by a sinusoidal curve whose amplitude decreases with the time or distance according to which case it represents. This is shown in Fig. 19, where the Y axis measures displacements, and the X axis measures time when the vibrating source is considered, and distance when the wave train is damped. In the first case the period T is a constant, being independent of the amplitude when the vibration is simple harmonic. In the second case, the wave length λ is also constant under similar conditions.

This fact is a very important one in the theory of damped vibrations, and seems at first sight paradoxical. It may be roughly explained by considering the case of a vibrating body whose amplitude decreases steadily with the values $a_1 > a_2 > a_3 > a_4 > \dots a_n$. Although in

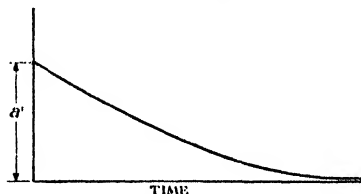


Fig. 20.

swinging from a_2 to a_3 it does not go as far as it did in the swing from a_1 to a_2 , and therefore would seem to require less time to go a shorter distance, its average velocity is also less, since the accelerations at the

end of each swing are proportional to the amplitudes, and thus slower speed compensates for the decreased distance.

If the damping is too great and exceeds a critical value, an initial displacement a' results in a gradual return to zero along a logarithmic curve asymptotic to the time axis, as shown in Fig. 20. This occurs when there is not sufficient kinetic energy to carry the vibrating mass over the point of rest to a displacement in the other direction.

SUPPLEMENTARY READING

J. A. Fleming, *Waves and Ripples*, Royal Institution Lecture, London, 1902.

J. W. Capstick, *Sound* (Chapters 1, 2, 3), Cambridge University Press, 1927.

PROBLEMS

1. It is 20 ft. from crest to trough of a train of water waves that pass a given point at the rate of 45 waves per min. What is their velocity in miles per hour? *Ans.* 20.5 mi./hr.

2. The displacement of a wave train 20 cm from its source and half a second after it started is -12 cm. The wave length is 60 cm, and the period of vibration is a quarter of a second. What is the amplitude? *Ans.* 13.9 cm.

3. A train of waves has a frequency of 20 v.p.s. Their amplitude is 8 cm, and their velocity 10 m per second. What is the displacement at a point 30 cm from their source one quarter of a second after the train started? *Ans.* 4.7 cm.

4. What are the velocity and the period of water waves 40 ft. between crests? *Ans.* 14.3 ft./sec.; 2.8 sec.

5. What are the velocity and frequency of ripples on mercury 2 mm between crests at 20°C ? (Consult tables in Articles 145 and 171.) *Ans.* $v = 32.8$ cm/sec.; $n = 164$ v.p.s.

CHAPTER 24

Sound and Its Transmission

332. Nature of sound. Sound may be defined in two ways—objectively and subjectively. Objective sound is a particular form of wave motion taking place in ponderable matter, whether gaseous, liquid, or solid, due to an original vibration or disturbance set up in the sounding body. Subjectively it is a sensory experience in the brain, conveyed to it from the ear. In this latter sense there is no sound when a tree falls in the forest, with no one to hear it. But in the objective meaning of the word there is always sound in such a case, regardless of its perception by any or no ear. Thus this ancient quibble, like most others, is seen to be wholly a question of definition.

It is objective sound which will be almost exclusively considered in the following chapters, for its subjective aspect belongs to the province of physiology and psychology.

333. Production of sound. Sound is invariably produced by motion of some sort set up in a gaseous, liquid, or solid body. If the motion is abrupt and discontinuous, the sound is sharp like the crack of a whip. If it continues, the sound may be harsh and unmusical like that produced by rubbing two stones together, or smooth and more or less pleasing, such as the tones of a musical instrument. The former is due to vibrations of no regular form or frequency; the latter to vibrations which are more or less defined, and have one or more definite frequencies. It is these periodic vibrations which are most interesting, and they will form the chief subject matter of what follows.

The vibrations which set up the sound waves may be of several possible types. The sounding body may vibrate transversely, longitudinally, or perform torsional vibrations by rapidly twisting back and forth around an axis. In any of these cases it must act upon some medium with which it is in contact, in order to set up in that medium the vibrations characteristic of sound. Without such a medium there is no sound, even in the objective sense of that word.

334. Propagation of sound. The medium which conveys sound must be a ponderable material. Sound cannot travel across empty

space, as is readily demonstrated by ringing a bell in an exhausted bell jar. The sound, which must travel through the air within the jar to reach the ear, becomes gradually fainter as that air is exhausted, and finally fades out almost entirely when a fairly high vacuum has been attained.

The fact that sound travels through gases gives us at once a clue to the nature of the motion which the sounding body sets up in the medium that carries the sound. Gases (and liquids except at or near their free surfaces) can vibrate only longitudinally, for transverse vibrations demand either rigidity (if they are to occur *within* a medium), or tension, as in a stretched string or surface film. Sound then consists of longitudinal vibrations of the conducting medium, and these are passed along from particle to particle, forming longitudinal waves such as have already been discussed. This mode of propagation has been established by innumerable experiments, some of which will be described farther on, so that no one need doubt the reality of sound waves.

Care should be taken to distinguish between the modes of production of sound and of its propagation. Any kind of vibration may be a source of sound, but only one kind constitutes sound itself and carries it from one point to another, and that kind is longitudinal vibration which results in a succession of condensations and rarefactions of the particles of the medium that carries it.

335. Reception of sound. Sound may be received not only by the ear, but by any device which absorbs vibrations and converts them into some other form of energy or motion. If absorbed by an inelastic substance like felt or sand, it must be converted into a minute quantity of heat, the lowest form of energy. It may however actuate a diaphragm equipped with a style which records the vibration on a moving disc, and thus it produces a phonographic record. The moving diaphragm may also cause variations in the electrical resistance of a telephone transmitter, thus setting up fluctuations in the current from a battery, and so originate corresponding vibrations at the receiving end of the line. These various receivers of sound waves convert the sound into some other kind of disturbance which is no longer sound, but may retain its characteristics translated into some other kind of vibration, such as the electromagnetic waves sent out by a broadcasting station. These are not sound at all, but are capable of re-creating sound waves at the receiving station.

336. The velocity of sound. The fact has long been known that sound takes a very perceptible time to reach the ear from a source

that is not too near, and is an experience familiar to everyone. Although light, as well as sound, has a finite velocity, it travels so fast that a light signal may be regarded as taking no appreciable time to travel the short distances over which sound may be heard. Therefore when we see a lightning flash and hear the thunder several seconds later, we infer correctly that the sound took just that time interval to reach us. When a distant locomotive whistles, we see the puff of steam before we hear the sound, and the report of a gun reaches us after the flash from the muzzle.

All the earlier attempts to measure the velocity of sound were based on timing this difference between the arrival of the sound and light signals of the same event, which had come over a sufficiently long distance to make it possible to measure this time interval with reasonable precision. The earliest of these attempts was made by Mersenne in 1640, who obtained a velocity of 448 m/sec. This is much too large, but only sixteen years later two Florentine academicians, Borelli and Viviani, obtained the surprisingly accurate value of 361 m/sec. Since then, with steadily improving methods, the results have shown a tendency to give an increasingly smaller velocity approaching the theoretical value based on calculation. This, as we shall see later, is 331.2 m/sec. at 0° C, and under standard atmospheric conditions.

The best experimental determination of the velocity of sound in free air was made at Sandy Hook by Dayton C. Miller, an American physicist, and the results were published in 1934. Coast-defense guns were the source of the sound, and the time of its arrival at a series of stations at measured distances up to four miles was accurately recorded. The receiver was a microphone like those used in broadcasting, and the electric current operated a "string" galvanometer whose deflections were recorded on a moving photographic film which registered the exact time the impulse arrived. The results gave a velocity of 1,087.13 ft./sec. at 0° C, or 331.36 m/sec., which is very close to the theoretical value.

Other determinations have been made by indirect methods, such as by Kundt's tube, to be described later, but these are chiefly valuable as methods for measuring the velocity of sound in gases other than air, or in liquids or solids. In any case they do not apply to the velocity of sound in *free* air.

337. Regnault's experiments. Regnault was the first to eliminate the "personal equation" of the observer of sound velocity. This he accomplished by recording automatically on a chronograph the

moment of firing a pistol, while the reception of the sound was also recorded when it impinged upon a flexible diaphragm. The firing of the pistol broke a fine wire which was part of an electric circuit reaching to the distant receiving station. This circuit contained an electromagnet which held a needle down upon the revolving drum of a chronograph, so that breaking the circuit released the needle and interrupted the record. The sound was received upon a flexible diaphragm, behind which was a very narrow gap between platinum contacts. The sound wave, acting on the diaphragm, closed this gap and completed a second circuit in parallel with the first. This again "made" the magnet, the needle was depressed, and the interrupted record was resumed. As the rate at which the chronograph drum revolved was known, the length of the interruption measured the time required for the sound to reach the diaphragm.

The chief interest in Regnault's experiments lies in his conclusions as to how certain conditions affect the velocity and intensity of sound in tunnels. He found:

(a) That sound diminishes in intensity as the distance it travels increases, and that this decrease was greatest in tunnels of small diameter.

(b) That the velocity of sound depends partly upon its intensity. Very loud sounds travel considerably faster than faint ones.

(c) That sound travels faster in large tunnels than in small ones, and apparently reaches a maximum in free air.

(d) That the velocity of sound is not affected by its mode of production. High-pitched sounds, for instance, travel as fast as those of low pitch. Conclusions (b) and (d) are especially significant. The greater speed of very loud sounds has occasionally been noticed, without any measuring device, when atmospheric conditions were such as to carry faint sounds far enough. In this way the report of a cannon has been heard before the command to fire.

The absence of any effect of the nature of the sound (except its loudness) is evident when we are listening to a band playing at a distance. The notes played simultaneously by the different instruments reach the ear together. If they did not, the effect would be most unpleasant, and indeed any music would be hopelessly distorted by distance.

338. Newton's formula. The theoretical velocity of a longitudinal wave in an elastic medium was derived by Newton, but as his method is rather difficult, the following simple proof is given in its place:

Since, as Regnault discovered, the nature of sound (really meaning

its wave form and frequency) does not influence velocity, we are at liberty to assume any kind of wave in calculating its speed. Let us then assume one whose compressions of constant pressure change abruptly to rarefactions of constant pressure also. The problem then really resolves itself into one of obtaining the velocity of a zone of compression which begins abruptly and then remains constant for its entire length. In Fig. 21 the shaded area represents the compressed medium having a specific volume v' . This compression (not

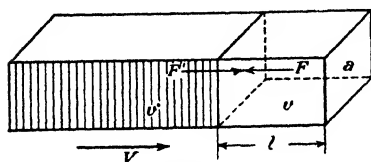


Fig. 21.

the medium) is moving to the right with a velocity V along a rod or tube of section area a . Let l represent the distance it will move in one second into the uncompressed medium whose specific volume is v . Evidently the distance l is numerically equal to

the velocity V . Now as the compression advances, the mass of each cubic centimeter increases, and this motion of increasing mass is the same thing as momentum. The total increase of mass per second is the increase within the volume la , but each cubic centimeter gains $d' - d = (1/v') - (1/v)$ grams, where d' and d are the densities. Therefore the total gain of mass per second is $la(1/v' - 1/v)$, and the time rate of change of momentum is

$$Vla(1/v' - 1/v) = V^2a(v - v')/vv', \quad (1)$$

since $V = l$ numerically. But the time rate of change of momentum is force, and the force in this case must be the difference between the total molecular forces F' and F acting at the boundary of the compression wave front, as shown in the diagram, where F' measures its forward push, and F the opposition of the undisturbed medium. Therefore

$$V^2a\left(\frac{v - v'}{vv'}\right) = F' - F. \quad (2)$$

Since pressure is force per unit area, we may divide by a giving

$$V^2\left(\frac{v - v'}{vv'}\right) = p' - p,$$

or

$$\frac{V^2}{v'} = \left(\frac{p' - p}{v - v'}\right)v. \quad (3)$$

By definition the right-hand member equals the elastic bulk modulus, B , of the medium. Then $V^2/v' = B$. But v' in the left-hand mem-

ber (the specific volume of the compressed medium) may now be set equal to the specific volume of the *undisturbed* medium. This is legitimate when changes of density are small, as is always the case when sound travels through liquids and solids, and usually so in air. Also, a compression is followed by a rarefaction where the value of v' is as much larger than v as it is smaller in a condensation, so the two approximations neutralize each other in calculating the velocity of the whole wave. Therefore $V^2 = Bv$, or

$$V = \sqrt{B/d}, \quad (4)$$

where d is the density of the undisturbed medium. This is Newton's formula, and it is applicable to all types of longitudinal waves in elastic homogeneous media, provided the proper modulus of elasticity is used.

339. Laplace's formula. We have seen in Article 257 that for small changes of volume the elastic modulus is approximately equal to the pressure provided Boyle's law applies. Therefore, making this assumption once more, we find that Newton's formula $V = \sqrt{B/d}$ becomes $V = \sqrt{p/d}$. If then this equation is used to obtain the velocity of sound in air, and the calculation is made for standard conditions, when $p = 1,011,330$ dynes/cm² and $d = 0.001293$ g/cm³, then $V = 280.26$ m/sec. This is obviously much too low, and was recognized as such by Newton himself. However, the difficulty was not explained until 1816, when Laplace pointed out that in deriving $p = B$, it was assumed that the temperature was constant and that Boyle's law was applicable. This is not the case, because the compressions of a sound wave heat the air, and the rarefactions cool it.

The reasons why these heatings and coolings do not neutralize each other, thus producing isothermal conditions, is explained as follows: In order to counteract the cooling of a rarefied region, heat from an adjoining region of compression must travel by conduction half a wave length during the time of half a vibration, $\lambda/2v$. From the theory of conduction we find that the *distance* through which an appreciable change in temperature is conducted varies as the square root of the time. It follows that the *velocity* of the conducted heat wave varies inversely as the square root of the time, so that the time required to travel half a wave length is longer with long waves and low frequencies than with short waves and high frequencies. Therefore, only in the case of *very* short waves could the process be isothermal. With ordinary sound waves it is adiabatic. Condon† has shown that

† E. U. Condon, *The American Physics Teacher* (Vol. 1, No. 1), February, 1933.

heat conduction becomes effective only when the wave length is comparable to the mean free path (10^{-5} cm) of the molecules of a gas. Thus the reason we have adiabatic rather than isothermal conditions is that the wave lengths of ordinary sounds are too great for effective conduction.

Under adiabatic conditions, $p\nu^\gamma$ is a constant, where γ is the ratio of the specific heats. Then, as proved in Article 257, $B_\phi = \gamma p$, where B_ϕ is the adiabatic modulus of elasticity. This is in contrast with B_θ , the isothermal modulus, which simply equals the pressure. If then we use B_ϕ instead of B_θ , Newton's equation reads

$$V = \sqrt{B_\phi/d} = \sqrt{\gamma p/d},$$

and since $\gamma = 1.40$ for air,

$$V = \sqrt{1.40p/d}.$$

With this important correction, the calculation of V agrees extremely well with the observed value. But in the case of most liquids and solids, the compressibility is so small that B_ϕ is practically equal to B_θ , which means that $\gamma = 1$ approximately, so that the distinction between the two moduli is unimportant.

340. Comparison of the velocities of sound in different media. Because of their high elastic moduli, metals and liquids conduct sound faster than gases, in spite of their greater densities which tend to lower V . Thus, whereas the adiabatic modulus of air at atmospheric pressure is given by $B_\phi = p\gamma = 1,011,330 \times 1.40 = 14.3 \times 10^5$ approximately, the volume modulus of copper is 14.3×10^{11} , which is 10^6 times larger. The density of copper is 8.9, which is about 6900 times as great as that of air, so that the reduction of V because of increased density is much more than offset by its high elasticity. The calculation of V for copper from $V = \sqrt{14.3 \times 10^{11}/8.9}$ gives a value of 4×10^5 cm/sec. very nearly, which is more than twelve times the velocity of sound in air.

In the case of water, we find its elastic modulus by taking the reciprocal of its compressibility. This is 48.9×10^{-6} per megabar (10^6 dynes per cm^2). Therefore $B = 10^6/(48.9 \times 10^{-6}) = 2 \times 10^{10}$ approximately. But the density of water is unity; therefore $V = \sqrt{2 \times 10^{10}} = 1.4 \times 10^5$ cm/sec. This is more than four times as fast as the velocity in air.

341. Effect of temperature on the velocity of sound. Since the density of the medium enters into both Newton's and Laplace's formulae, temperature must affect the velocity of sound in all media. The expansion caused by heating means a decrease of density, and

a consequent increase of velocity. But in solids and liquids this effect is so small that it may usually be neglected. In air and other gases however it is very important.

The relation due to Charles, $v_t = v_0(1 + \alpha t)$, may be converted into one between densities, because $d = 1/v$. Therefore $d_t = d_0/(1 + \alpha t)$. Then the velocity of sound, $V_t = \sqrt{\gamma p/d_t}$, in a medium whose temperature is t and density d_t , may be expressed in terms of the standard density, d_0 , by substituting for d_t in the velocity equation, so that

$$V_t = \sqrt{\frac{\gamma p(1 + \alpha t)}{d_0}} = \sqrt{\frac{\gamma p}{d_0}} \times \sqrt{1 + \alpha t}.$$

But $\sqrt{\gamma p/d_0} = V_0$; therefore

$$V_t = V_0 \sqrt{1 + \alpha t},$$

and this is approximately equal to

$$V_t = V_0(1 + \frac{1}{2}\alpha t)$$

for temperatures not too far from 0° C. Thus we may calculate the velocity at any temperature if α and the velocity at 0° C are known.

In the case of air, $\alpha = 0.003665$, and if we set $t = 1^\circ$ C, and take $V_0 = 331.36$ meters per second, $V_t = 331.967$, which shows an increase of 60.7 centimeters per second per degree rise of temperature in air.

342. Velocity of loud sounds. The effect of the atmospheric pressure on sound velocity is zero, because, though it appears in the equation $V = \sqrt{\gamma p/d}$, the density varies directly as the pressure, and the ratio $p:d$ is constant at constant temperature. But the rapid changes in pressure, owing to the condensations and rarefactions of a sound wave, may influence V , if they are large enough. In proving $B_t = p$ it is assumed that these changes are small compared to the pressure of the atmosphere, but if they are not, and Δp is relatively large, then $p + \Delta p$ must be used instead of p to replace the elastic modulus in the equations for the velocity. Newton's equation then reads $V = \sqrt{(p + \Delta p)/d}$, or if we introduce Laplace's correction, $V = \sqrt{\gamma(p + \Delta p)/d}$.

When very large guns are fired, Δp may be actually as large or even larger than p , which would mean compressions of two or more atmospheres, diminishing of course with the distance. At any point where $\Delta p = p$, the velocity would be $\sqrt{2}$ times as great as it would be for faint sounds where Δp is negligible.

343. Sound ranging. The position of an enemy gun in the World War was frequently located with remarkable precision by the use of

a range finder based on sound instead of light. This was accomplished by recording the exact time at which the report of the gun reached three receiving stations located at some distance from each other. Thus in Fig. 22 the gun is located at P , with a receiving station at A , at B , and at C , where by means of microphones and chronographs

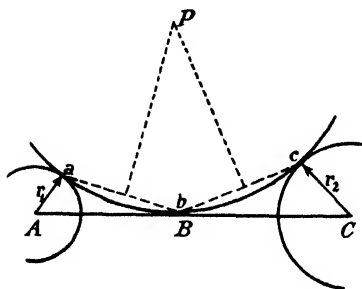


Fig. 22.

the time of the sound's arrival is recorded. Suppose that the time of arrival at A is t_1 seconds after it reaches B , and that it reaches C t_2 seconds later than B . Then draw circles around A and C whose radii, r_1 and r_2 , are equal to the distance the sound travels in the times t_1 and t_2 . That is, $r_1 = Vt_1$, and $r_2 = Vt_2$, where V is the velocity of sound.

The circular wave front abc must be at a distance r_1 from A and r_2 from C at the moment it reaches B . Consequently abc is the arc of a circle drawn tangent to the two smaller circles and passing through B . The gun P , which is obviously at the center of this arc, is therefore located.

SUPPLEMENTARY READING

- J. W. Capstick, *Sound* (Chapters 4, 5), Cambridge University Press, 1927.
 F. R. Watson, *Sound* (Chap. 17), Wiley, 1935.
 A. B. Wood, *A Textbook of Sound* (Section 3), Macmillan, 1932.
 E. G. Richardson, *Sound* (Chap. 1), Arnold, London, 1927.

PROBLEMS

1. Calculate the velocity of sound in steel (density 7.8 g/cm^3) from Newton's formula. *Ans.* 4822.2 m/sec.
2. Calculate the velocity of sound in oxygen at 0°C and normal atmospheric pressure. ($\gamma = 1.40$). *Ans.* 315.1 m/sec.
3. What is the exact velocity of sound in air at a temperature of 30° , taking $V_0 = 332 \text{ m/sec.}$? *Ans.* 349.8 m/sec.
4. The flash of a gun precedes the report by 8 sec. in air at 59°F . Taking the velocity of sound at 32°F as 1210 ft./sec. , what is the actual velocity, and how far off is the cannon? *Ans.* 1243 ft./sec.; 9944 ft.
5. The velocity of sound is found to be 1385 m/sec. in mercury. What is its modulus of volume elasticity? *Ans.* $2.6 \times 10^{11} \text{ dynes/cm}^2$.

CHAPTER 25

Properties of Sound

344. How sounds differ. Sounds differ fundamentally in three ways. They differ in *intensity* (and consequently in loudness), in *pitch*, and in *timbre*, or "tone color." Intensity depends upon the energy contained in unit volume of the medium at any instant and is defined as the *time rate of flow of energy across a square centimeter at right angles to the direction of propagation*. In the case of a plane wave, this is equal to energy density times the velocity of the sound. Pitch depends upon the frequency of vibration of the sounding body, that is, on *the number of vibrations it executes per second*. And timbre, in a complex tone like that of the violin, depends upon the *resultant wave due to a combination of frequencies*.

345. Intensity. The energy of any harmonic vibration, and therefore the intensity of sound, varies as the square of the amplitude and the square of the frequency. This may be proved as follows: If a mass m is executing simple harmonic vibrations, its velocity at any instant may be found (equation (2) Article 93), from $v = -\omega r(\sin \omega t)$, where r is the amplitude. This is a maximum when $t = T/4$ or $3T/4$, for then $\sin \omega t = 1$. Therefore $v_m = -\omega r$.

Now such a mass has its maximum kinetic energy when its velocity is a maximum, and this occurs in mid-swing. At this point the potential energy is zero, and therefore the kinetic energy represents the total amount W . Substituting for v_m in $W = mv^2/2$, we obtain

$$W = m\omega^2 r^2/2. \quad (1)$$

Then since $m = du$, where d is the density of the undisturbed medium, and u is the volume occupied by m , the energy per unit volume is given by

$$W/u = \omega^2 r^2 d/2, \quad (2)$$

and the rate of flow across unit cross section, or intensity, is

$$WV/u = I = \omega^2 r^2 Vd/2, \quad (3)$$

when V is the velocity of sound. Substituting $\omega = 2\pi n$, we have

$$I = 2\pi^2 n^2 r^2 V d. \quad (4)$$

Since then m or d are supposed constant, *the intensity of a given sound varies directly as the squares of the amplitude and of the frequency.*

It may also be shown that the intensity is given by

$$I = \frac{P^2}{2Vd}, \quad \text{or} \quad \frac{p^2}{Vd}, \quad (5)$$

where P is the maximum value of the increase in pressure due to the sound, or "excess pressure," and p , which equals $P/\sqrt{2}$, is its "effective," or square-root-of-mean-square, value, denoted by r.m.s. Equation (5) is convenient in calculating the sound produced by a vibrating diaphragm which produces the compression, as in the telephone receiver, phonograph, and so forth.

The intensity of sound of a given frequency and in a given medium, which determines d and V (equation (4)), varies only with the square of the amplitude. But this depends upon three factors. The first is the amplitude of vibration of the source of the sound, which is sufficiently obvious. The second is the area of the sounding body, which has the same influence upon the intensity at a distance as the area of a heated body has upon the intensity of heat radiation at a distance. The third is the fact that the intensity varies inversely as the square of the distance.

The influence of area cannot easily be proved, but the general principle is demonstrated by placing the shank of a vibrating tuning fork upon a table top. The vibrations, rather faint before, are now quite loud, owing to the greater area set in vibration. The fork, however, stops sounding much sooner than if held in the hand, because its energy is partly communicated to the table, thus producing more sound, but for a shorter time.

The inverse square law, which applies rigorously only to a point or spherical source and a homogeneous medium, has already been derived for the general case of radiant energy, and it is therefore not necessary to prove it for the particular case of longitudinal vibrations. If the source is not a point, but a vibrating string, for instance, the law is far from true near the string, but becomes increasingly exact at increasing distances, when the dimensions of the vibrating object become small in comparison.

346. Pitch. The frequency of vibration determines whether a sound is pitched high or low. A high pitched sound is said to be *acute*, and a low pitched sound, *grave*.

The fact that the pitch of sound depends upon the frequency is easily demonstrated by holding a card against the teeth of a rapidly rotating gear wheel. The resulting tone rises in pitch as the wheel speeds up. Another method is to direct an air blast through a nozzle against a rotating disc in which equally spaced holes have been punched, as shown in the outer circle of Fig. 23. If the blast is directed against the inner row of unequally spaced holes, the sound has no definite pitch, and is in the nature of an unmelodious roar.

The human ear is limited in the range of frequencies which it can perceive as sound. Investigations in this field conducted by the great German physicist Hermann von Helmholtz (1821-1894) led him to conclude that sounds of fre-

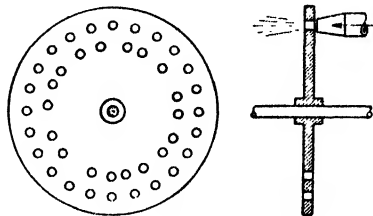


Fig. 23.

quencies lower than 16 to 20 vibrations per second (v.p.s.) could not be perceived as a definite tone, but merely as a throbbing sensation in the ear. The upper limit of audibility by the human ear is between 20,000 and 25,000 v.p.s., while many persons of normal hearing at ordinary frequencies, cannot perceive sounds higher than 10,000 v.p.s.

Very acute tones cannot be clearly distinguished from each other as regards pitch; therefore the range available for an orchestra is limited to tones below about 5000 v.p.s. Actually the highest note commonly employed in orchestral music (disregarding overtones) is the high *d* of the piccolo, whose frequency is 4702 v.p.s., while the lowest is the *E* string of the double bass with 41 v.p.s. This range lies between wave lengths of 8.4 meters at the bottom and 7.3 centimeters at the top.

347. Timbre. The peculiar quality of a sound, which enables us to distinguish between different sources giving the same pitch, as between the *A* of a violin and of a flute, depends upon the *overtones* present. Overtones are vibrations of a frequency higher than that of the principal tone, or *fundamental*, upon which they are imposed by complex vibrations of the sounding body. These overtones are harmonic vibrations and the quality or timbre of the tone depends upon their frequencies and amplitudes.

The earliest careful investigation of the causes of timbre was by von Helmholtz, who made a qualitative analysis of complex tones. This was done by means of resonators, each of which responded to

but one frequency. He also effected the synthesis of complex tones out of simple harmonic vibrations by means of a kind of organ made of electrically driven tuning forks whose pitches were the harmonics of a single fundamental. Different combinations of these produced different tone colors.

More recently Professor Miller has analyzed complex sound vibrations by means of an ingenious device which he named the *phonodeik* (sound exhibitor). The essential parts of this instrument are a horn

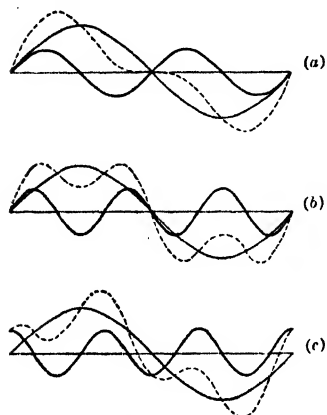


Fig. 24.

and diaphragm to receive the sound, and a small mirror which rocks upon a horizontal axis and rotates back and forth as a result of the diaphragm's vibrations. A beam of light reflected from the mirror then moves in a vertical plane upon a moving photographic film drawn horizontally. The resulting "waves" traced upon the film are similar to the dotted curves shown in Fig. 24, though more complicated, and they may be analyzed into their simple harmonic components. The frequency and amplitude of these components are found, and it is thus possible to deter-

mine the "ingredients" which give the tones of a violin or a flute, or any other instrument, their characteristic tone colors. The wave forms of the different vowels are also separated into their components and the voices of singers analyzed, as well as many other kinds of sound.†

348. Complex periodic vibrations. If a complex tone is to be strictly periodic and suitable for music, that is, having a definite wave form or series of wave forms which are always identically repeated, the frequencies of the overtones must be definite multiples of some fundamental frequency.

In the usual case, where the waves all have the same form, the overtones are integral multiples of the fundamental. If its frequency is n , then the lowest possible *overtone* (or partial) has a frequency of $2n$, and is known as the second *harmonic*. The next possible overtone, $3n$, is the third harmonic, and so on. If the complex tone contains the fundamental, or first harmonic, of frequency n , and the

† The most modern and accurate method of analyzing sound-wave forms involves the use of a microphone and analysis of the electric currents set up by the motion of the diaphragm caused by the sound waves acting upon it.



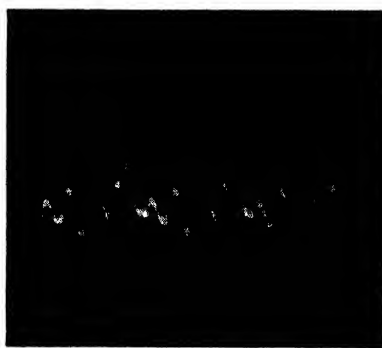
(a)
Middle F of flute



(b)
Middle F of flute



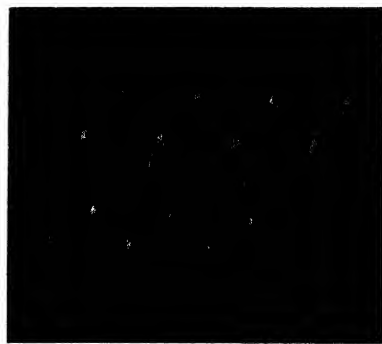
(c)
Low E of clarinet



(d)
Middle B of clarinet



(e)
Low G of oboe



(f)
Low F of horn (melophone)

Plate 3.

Photographs of "wave forms" made with condenser microphone and oscillograph. Both (a) and (b) show weak second harmonics, but as only the phase has shifted, both sound alike. Waves (c) and (d) show that low tones of a wind instrument are richer in harmonics than those in the upper registers. Wave (f) exhibits a very strong second harmonic slightly out of phase with the fundamental.

odd overtones, $3n$, $5n$, and so forth, the third harmonic, $3n$, is the first overtone. But if both even and odd harmonics are present, $3n$ is the second overtone.

The perception of timbre is due to an analysis of the complex tone by the ear, which breaks it down into its component vibrations. These are then transmitted to the brain, where a partial synthesis is effected, giving us the sense perception of the tone as a whole. The resulting sensation therefore depends upon the particular overtones present and their relative amplitudes, but not upon their phase relations.

349. Graphical summation of harmonics. Although the quality of a complex tone as perceived by the ear appears to be independent of

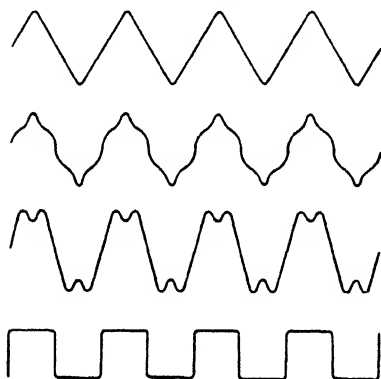


Fig. 25.

phase relations, these do alter the resulting form of the vibration when depicted as a transverse wave. In Fig. 24 (a) the fundamental and second harmonic are shown in phase with each other, but having different amplitudes. Their resultant is indicated by the dotted line. The second loop of the resultant is like the first, except that it is both inverted (upside down) and perverted (turned left for right). In (b) is seen the effect of a third harmonic in phase

with the fundamental, and in (c) the third harmonic is 90° out of phase. In these two cases the second loop is only inverted, for in (c) it is evident that right and left have not been interchanged. This is a general property of all *odd* harmonics, because, as is shown in (b) and (c), the component sine curves are identically related in two successive half periods when reversed, while in (a) they are not, and *even* harmonics therefore produce perversion of the wave form.

By combining with the fundamental enough harmonics having the requisite amplitudes, frequencies, and phase relations, any conceivable wave form may be produced. A few such wave forms are shown in Fig. 25. The lower one with almost perfectly rectangular teeth would require one hundred or more harmonics properly arranged to produce it.

350. Reflection of sound. We have already seen how a longitudinal wave is reflected at an interface between two media of different

density. The most striking case of this kind of reflection is the echo. In order to have the reflected sound sufficiently separated from the original to distinguish it, the source must be at a considerable distance from the reflecting surface, otherwise the two will merge into a confused jumble.

The echo from a picket fence stretching beyond an observer may be made to yield a tone of some duration and definite pitch as a result of a single sharp sound, such as that produced by clapping the hands together. The successive echoes from individual pickets arrive each a little later than the other, and are nearly equally spaced in time, thus producing the effect of a definite vibration frequency. The prolonged roll of distant thunder is due to repeated reflections of the original "clap" from layers of air of different densities and at different distances from the observer.

Reflected sound may be both an advantage and a disadvantage in churches, lecture halls, and theaters. In many churches a sounding board is placed behind the pulpit for two reasons. The part of the sound of the preacher's voice which would otherwise travel away from the congregation is caught and reflected toward them, together with those waves which started forward in the first place. These two beams of sound are slightly out of step, but not enough so to produce confusion. Without the sounding board, the reflection would occur from walls and other surfaces farther back. If the reflecting surface is a smooth wall which does not "break up" the sound as pillars and arches would, the echo may be quite distinct from the original sound, and as it has farther to go, the two sounds may be so much out of step as to cause serious confusion. This depends upon the distance between the speaker and the reflecting surface, which should not exceed twenty-five feet. There would then be a fifty-foot difference of path between the sound and its echo, and this results in a retardation of about one twentieth of a second, which is the longest interval permissible if the separate syllables of speech are not to overlap in a confusing manner. Thus a sounding board serves the double purpose of amplifying the sound by increasing the energy transmitted forward and of cutting off an undesirable echo.

In halls of complicated form, another source of trouble is the tendency for the reflected waves to concentrate at some points or regions known as *foci*, while other regions are left almost void of sound. This may be illustrated experimentally by two parabolic reflectors with the source of sound in the focus of one of them and the ear at the focus of the other. It is a well-known property of the parabola that a line

drawn parallel to its axis to meet the curve at the point p (Fig. 26) makes the same angle with the normal, n , to that point as a line drawn between p and the focus s . Thus the angles i and r are equal, so that all sound (or light) rays proceeding from the focus are reflected as rays parallel to the axis. The second parabolic reflector receives this beam and focuses it in the reverse manner on the focus E . Therefore a very feeble sound like the ticking of a watch may be heard at E ,

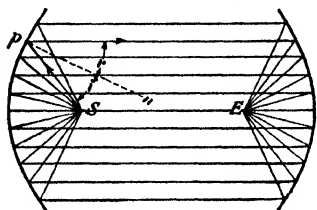


Fig. 26.

when the reflectors are so far apart that the sound would otherwise be quite inaudible.

351. Refraction of sound. Like all other waves, the front of a sound wave is altered, in shape or direction or both, when it enters a medium where it travels at a speed different from that which it had before. This may be shown by making a lens of thin sheet rubber filled with the heavy gas carbon dioxide, in which sound travels more slowly than in air. If sound, from such a distance that its wave front is almost plane, strikes such a lens, the center of the beam whose diameter is d (Fig. 27) reaches the lens first and is slowed down as indicated by the dotted lines. This retarding process progressively deforms the plane front into a curved one, concave forward. Then when these wave fronts begin to emerge, their exterior zones come out first and speed up, while the central portion which gets out last is still further retarded, thus increasing the concavity. Finally these

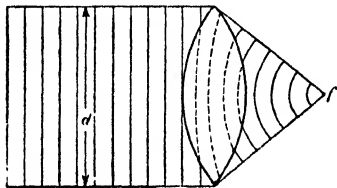


Fig. 27.

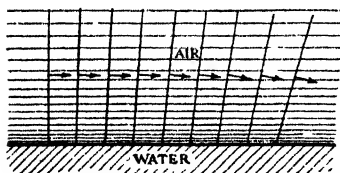
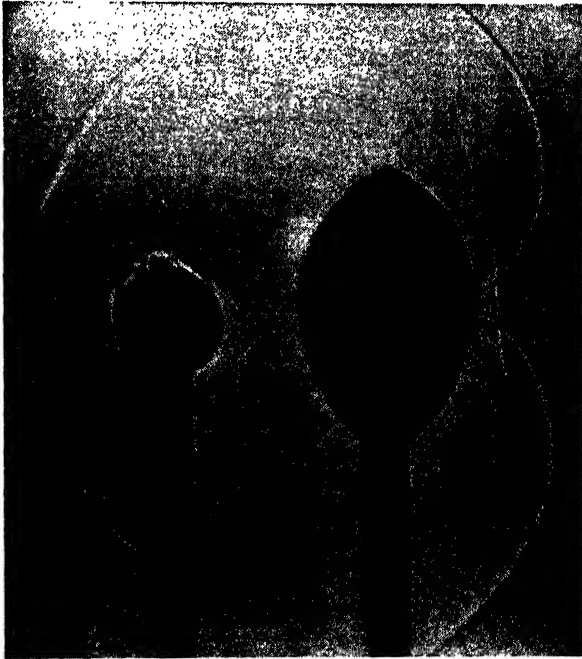


Fig. 28.

concave wave fronts converge at a point, or *focus*, f , where the sound is much more intense than it would be but for the lens.

As sound travels faster in warm air than cold, the different layers of air over the earth's surface cause a bending of a plane wave front such as has already been discussed. The air near a cold body of water is often cooler than that above. Then the upper strata are progressively less dense, and sound moving parallel to the water's surface at

higher levels travels more rapidly than it does down below in air of greater density. This results in a bending forward of the wave front, as indicated in Fig. 28, and the sound tends to cling, as it were, to the surface of the water, so that it may be heard a long way off. If, on the other hand, the air is warmer below than above, as over a



Courtesy Professor A. L. Foley, Indiana University

Plate 4.

Photograph of sound wave passing through a lens filled with sulphur dioxide gas. The circular wave near the left of the source is the wave front reflected from the lens. The double curve on the right is the original wave front diffracted around the lens. The short curve concave outward is the converging wave refracted by the lens.

desert, the reverse effect takes place. The wave fronts are tilted backward as they advance, and the sound tends to rise above the earth's surface.

The same effect may be observed when a gentle breeze is blowing over the earth. Near the surface it moves more slowly than it does higher up, because of the earth's roughness and resulting friction. Thus if the sound travels with the breeze, its wave front bends forward, as in Fig. 28, and so keeps to the ground and carries farther. Sound traveling against the breeze is more retarded by the more

rapidly moving air at the higher levels. Its wave front bends backward, carrying it away from the ground, and it becomes inaudible at a short distance from the source. It is thus clear that wind does not blow the sound to or from us to any great extent, for even a hurricane traveling at 100 miles an hour is moving slowly compared to sound, which goes 742 miles an hour. So sound is certainly not much slowed down or speeded up by an ordinary wind.

352. Interference of sound. Sound waves interfere in the same way as water waves already described, though of course in this case

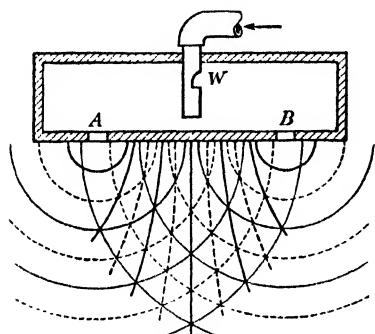


Fig. 29.

condensation and rarefaction take the place of crest and trough. In both cases when waves from two exactly similar sources combine, they produce hyperbolic loci of minimum and maximum intensity, or nodal and antinodal lines. This may be shown by enclosing a source of sound, such as a whistle, in a box from which the sound issues through two holes, the common source insuring the identity of frequencies which is

absolutely necessary. Such a device is illustrated in Fig. 29 with a part of the wave diagram. According to Huygens' principle, the two holes may be regarded as independent sources of sound, and as they are equidistant from the whistle *W*, the waves issuing from them are in the same phase, although that is necessary only if a symmetrical pattern of nodal and antinodal curves is desired.

353. Beats. If two notes of equal intensity but of slightly different pitch are sounded together, the resultant sound produces a succession of pulsations known as **beats**, which are due to the two wave trains

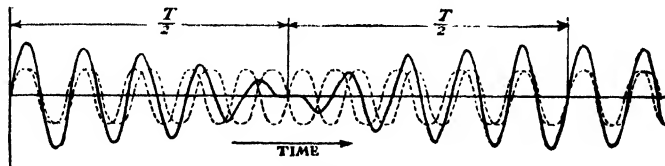


Fig. 30.

getting into and out of step at regular intervals. This may be understood from Fig. 30, where the two dotted curves represent the vibrations of the two tones reaching the ear. One of them, *A*, has a shorter

wave length than the other one, B , and drops behind it in phase, until after a number of vibrations it has lost half a wave length, and the two are then in opposite phase at the end of half a period of the beats. Half a period later, B has lost an entire vibration, is again in phase with A , and the two sounds reinforce each other as at the beginning. The resultant, or sum, of the two vibrations is indicated by the heavy line. Their decreasing amplitude followed by a corresponding increase represents the *beating* of the two notes.

If the frequency of A is n_1 vibrations per second, and if n_2 is the frequency of B , then A gains $n_1 - n_2$ vibrations over B in one second. Each time it gains a complete vibration, the two sounds are in the same phase and reinforce each other. There are then $n_1 - n_2$ maxima per second, or the frequency of the beats equals the difference of the frequencies of the two notes.

When the notes are nearly in tune, the beats are slow, but as the difference of frequency increases, the beats become increasingly rapid and may ultimately merge into a continuous tone known as a **difference tone**, which is much lower than either of the two which produced it. This phenomenon is made use of in certain organ stops which give very grave tones by using pairs of relatively short pipes of slightly different pitch.

354. Doppler's principle—Moving source. If the distance between the ear and the sounding body changes at a constant rate, the ear perceives a different pitch from that which it would with both at rest. The whistle of an approaching locomotive sounds more shrill than if it were at rest, and its pitch falls after it passes the observer. There are two cases of Doppler's principle, first formulated in 1842 by Christian Doppler of Prague. These are: Case I, when the sounding body moves toward or away from the ear; and Case II, when the ear moves toward or away from the sound.

Let us suppose the sounding body to be moving toward the ear with a velocity v . Then each successive compression it emits is nearer the preceding one than it would be if it were at rest. This decrease in the wave length is equal to the velocity v divided by the frequency; therefore the modified wave length λ' is v/n shorter than it was before the motion began; hence

$$\lambda' = \lambda - \frac{v}{n}. \quad (1)$$

But the frequency of these shortened waves must be correspondingly raised according to the relation $n' = V/\lambda'$, where V is the velocity

of the sound. Therefore, substituting for λ' its value from equation (1), we obtain

$$n' = \frac{V}{\lambda - v/n} = \frac{Vn}{\lambda n - v}.$$

But $\lambda n = V$; $\therefore n' = \left(\frac{V}{V - v} \right) n,$ (2)

where $V/(V - v)$ is the correction factor by which this case of the Doppler effect is to be calculated. Since the correction factor is greater than unity, n' is greater than n , which means that a note emitted by a body approaching the ear sounds more acute than if it were at rest, and the pitch rises with increasing velocity of the source.

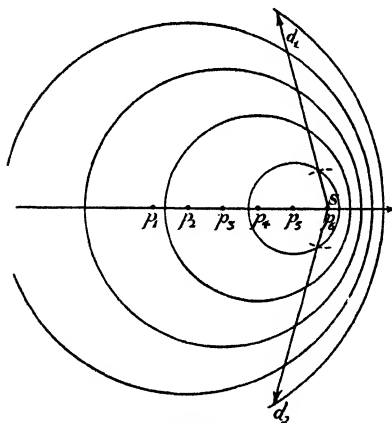


Fig. 31.

But if the source of sound were receding, v is negative and the factor becomes $V/(V + v)$, which is less than unity and shows that the pitch of the note perceived by the ear is lowered.

Both possibilities of Case I are shown in Fig. 31, where a moving body S is sending out spherical waves. Owing to its motion they are seen to be crowded together in front of it, and lengthened behind. In other directions the wave length has intermediate values, and in the directions d_1 and d_2 , λ is the same as if the body

were not moving. The circles represent successive wave fronts sent out when S was at the equally spaced points p_1, p_2, p_3, p_4 , and so forth. Therefore their radii differ by the same amount. This difference is the distance sound travels while the source moves between two adjacent points, and is equal to the radius of the circle around p_5 .

355. Doppler's principle—Moving observer. If the ear moves instead of the source, the case is different. Here there is no change in λ , but the ear encounters the waves at a different rate. Let v be the velocity of an ear approaching the source; then the increase in the number of waves received per second is obviously v/λ . Therefore the total number n' received per second is $n + v/\lambda$. Then

$$n' = n + \frac{v}{\lambda}, \quad (1)$$

and
$$n' = \frac{n\lambda + v}{\lambda}. \quad (2)$$

Multiplying numerator and denominator by n , and setting $n\lambda = V$, we obtain

$$n' = \left(\frac{V + v}{V} \right) n. \quad (3)$$

Here the Doppler effect is calculated by the factor $(V + v)/V$, which is not the same as $V/(V - v)$, although both are greater than unity and are nearly equal if v is small compared to V . Thus if $V = 100$ and $v = 1$, $(V + v)/V = 101/100$, while $V/(V - v) = 100/99$. But the first value is less than the second; hence the Doppler effect is a little greater when the source is approaching the ear than when the ear is approaching the source with the same velocity.

In Case II, if the ear is receding from the source, v is again to be regarded as negative and the correction factor is $(V - v)/V$, which in turn may be compared with $V/(V + v)$, the factor for a receding source. These also will be seen to be slightly different.

The difference between the two cases of the Doppler effect is due to the fact that the air furnishes a "frame" to which the motion may be referred. Therefore motion between the two objects is not a relative affair in such a case, and fixity with respect to the air has a real meaning. But in a vacuum, as in interstellar space, if we assume it could carry sound, this distinction would not exist.

Any kind of wave motion has a Doppler effect, but in the case of light, whose velocity through space cannot be referred to any fixed medium of reference, there is no difference between the cases of a moving source and moving observer. In either case the motion is wholly relative.

SUPPLEMENTARY READING

J. W. Capstick, *Sound* (Chapters 6, 7, 8), Cambridge University Press, 1927.

F. R. Watson, *Sound* (Chapters 6, 7, 8, 9), Wiley, 1935.

D. C. Miller, *The Science of Musical Sounds*, Macmillan, 1922.

PROBLEMS

1. Calculate the flow of energy across a square centimeter of a very loud sound in air, whose frequency is 256 v.p.s. and whose amplitude is 5 mm when the temperature is 0°C . *Ans.* 1.39 watts per cm^2 .

2. The whistle of an approaching locomotive has an actual pitch of 320 v.p.s., but its apparent pitch is 360 v.p.s. How fast is the locomotive going? *Ans.* 36.9 m/sec.

3. If the whistle in Problem 2 is at rest and the observer is approaching it in an automobile at the rate of 36.9 m/sec., what would be the apparent pitch of the sound? *Ans.* 355.6 v.p.s.

CHAPTER 26

Hearing and Acoustics

356. The human ear. Our hearing mechanism is extremely complicated and not fully understood. The ear is made up of three principal parts—the external, middle, and inner ear. In Fig. 32 is shown its structure much exaggerated, and with part of the bony covering of the inner ear cut away to lay bare its convolutions.

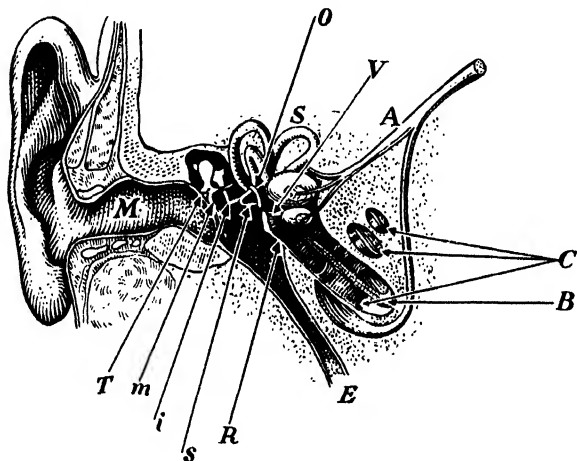


Fig. 32.

Sound waves falling on the visible part of the outer ear pass through the auditory canal *M* and cause the drum *T* to vibrate. These vibrations set in motion a series of little bones—the hammer, anvil, and stirrup, *m*, *i*, and *s*—which pass along the vibrations across the middle ear to a membrane, *O*, called the *oval window*, communicating with the *labyrinth*, or inner ear. The middle ear connects with the mouth by means of the Eustachian tube, *E*, which serves to equalize the pressure on each side of the drum. Its lower end is closed by a valve which opens when we swallow, so that rapidly changing pressure on the drum, as experienced in a moving elevator, with a consequent “bubble in the ear,” is equalized by swallowing.

The inner ear consists of three parts—the semicircular canals, *S*, the vestibule, *V*, and the cochlea, *C*. It is filled with a watery fluid

known as *endolymph*. The semicircular canals are connected with the brain by a branch of the auditory nerve, *A*, and are the source of our sense of equilibrium. These canals lie in three planes—one horizontal, and two vertical at right angles to each other. They thus represent the three spatial dimensions in which they act somewhat like spirit levels.

The vestibule is equipped with many tiny elastic hairs projecting from its inner walls, and when the lymph is set in motion from the oval window, they respond and probably serve to give us a sense of sound in general, but devoid of tonal quality, such as scraping or hissing sounds.

As the endolymph, like all liquids, is highly incompressible, it can be set in vibration only if it is in contact with some body more flexible than the bony cavity surrounding the labyrinth. This need is supplied by *R*, the *round window*, which is closed by a membrane similar to that of the oval window, and this bulges out when the oval window bulges in, and vice versa.

The *cochlea* is a spiral tube of about two and a half turns, resembling a snail shell, from which it is named. It is separated for nearly its entire length into three long galleries, *v*, *c*, and *t*, as shown in Fig. 33.

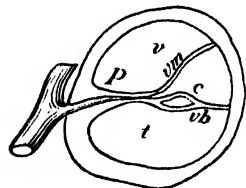


Fig. 33.

These are formed by a bony partition, *P*, reaching part way across, to which are attached two membranes stretched out to the opposite wall of the cochlea. Nerve fibers pass through the bony partition to the lower membrane, *vb*, indicated also by *B* in Fig. 32. The upper gallery ends in the vestibule, and the lower gallery at the round window. Thus these two diaphragms are separated by the membranes except at the apex of the cochlea, where there is an opening, the *helicotrema*, which connects the galleries.

The membrane *vb* is called the *basilar membrane*. It increases in width from about 0.04 mm at the base of the spiral to nearly 0.5 mm at its apex, and is composed of thousands of transverse fibers stretched tightly across it like the strings of a harp. These fibers are supposed to respond to different pitches when set in vibration by the vibrating lymph; and as the longest are near the apex and farthest from the round window, it is by these that low tones are perceived, while high-pitched notes act upon the shorter fibers near the base of the cochlea. This is in accordance also with the hydraulic features of the mechanism, for on account of the inertia of the fluid, high-frequency

fluctuations would not easily be transmitted through its entire length. Thus high-pitched vibrations are transmitted across the membranes from the upper to the lower gallery, leaving the lymph higher up in the cochlea undisturbed. But notes of low pitch set the entire mass of the fluid into vibrations which are passed along from one gallery to the other by way of the helicotrema at the upper end of the spiral where the longer fibers are located.

Mounted on the basilar membrane are about 3000 pairs of fibers, or "rods," indicated at *c* in Fig. 33. These are known as the *arches of Corti*. Their function is not clear, but they may serve to pick up the vibrations of the endolymph and communicate it to the fibers of the basilar membrane, or they may simply help to tune the fibers which they partially span.

357. Loudness. This is a psychological experience, and depends upon the mechanism of the ear and brain. It increases with the intensity of the vibrations received, but there is no *simple* law expressing the relation between loudness and intensity. However, there is an equation based on a principle known as *Fechner's law* which makes it possible to calculate *S*, the sensation of loudness. This is given in terms of the maximum value *P* of the *excess pressure* of the sound wave, which means the maximum increase in pressure in a condensation over and above the pressure of the undisturbed medium. The equation is

$$S = c \log_{10} P + a,$$

where *c* and *a* are constants depending on the frequency, though *a* is also a function of the "threshold pressure" to be defined. Its value is small compared to $c \log_{10} P$, when the sound is loud, so that we may compare loud sounds by comparing the logarithms of their excess pressures.

Comparing the loudness of two sounds by means of the logarithms of their excess pressures amounts to comparing the logarithms of their intensities, because since $I \propto P^2$ (equation (5) Article 345), $\log I \propto 2 \log P$, or $\log I \propto \log P$. The difference, *B*, between two loudnesses is then measured by the difference between the logarithms of their intensities. That is, $B = \log I_1 - \log I_2$, or $B = \log (I_1/I_2)$. If $I_1 = 10I_2$, and if the logarithm is taken to the base of 10, then $B = 1$. This unit difference is known as the *bel*. So, measured in bels, $B = \log_{10} (I_1/I_2)$.

The bel is obviously a large unit, and the more convenient *decibel* (one tenth of a bel) is preferred, especially as it measures about the

minimum difference in loudness which the ear can distinguish. Then, measured in decibels, $b = 10 \log_{10} (I_1/I_2)$.

When b is one decibel, $I_1/I_2 = \log^{-1} 0.1 = 1.259$. This means that if one sound is a decibel louder than another, the intensity of one is 26 per cent greater than that of the other, a difference barely observable by the ear. If the sounds differ by three decibels, the intensity ratio is 1.995, or nearly two to one. If they differ by twenty decibels, $I_1 = 100 I_2$, and thirty decibels means that $I_1 = 1000 I_2$.

In order to be audible at all, a sound must exceed a certain minimum intensity. In recent investigations conducted in the Bell Telephone Laboratories, the intensities required to pass the threshold of audible

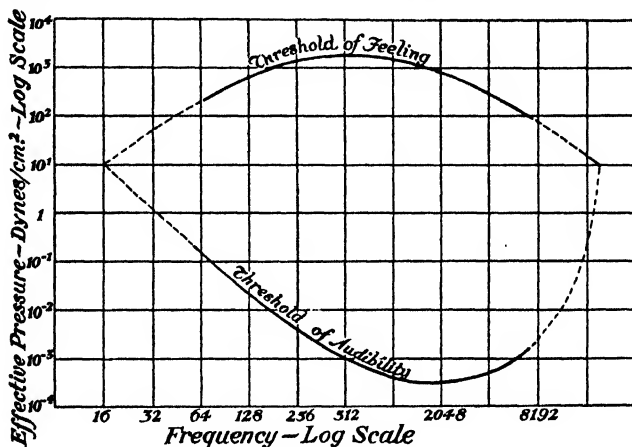


Fig. 34.

perception have been determined by measuring the effective value p_0 of the excess threshold pressure and then calculating I_0 from equation (5) cited above. The value of p_0 falls rapidly with rising frequency, from 0.12 dyne/cm², when the frequency is 64 v.p.s., to 0.00042 dyne/cm², with 4096 v.p.s.

For purposes of comparison of sounds of different pitch, the threshold pressure is taken as 0.001 dyne/cm², or one millibar. This corresponds to a sound intensity of about 24×10^{-16} watt/cm² under standard atmospheric pressure at 20° C, as is readily calculated from equation (5) referred to above.

The curves of Fig. 34, due to Wegel of the Bell Telephone Laboratories, show both the upper and lower limits of the "auditory sensation area." The effective (r.m.s.) value of the pressure changes which

affect the ear as sound are plotted against frequency, using a logarithmic scale. From an inspection of the lower, or threshold curve, it is evident that the pressure of one millibar assumed above corresponds to a frequency of the note c'' , or 512 v.p.s., and is close to the middle of the normal range of audibility. Then taking this frequency, we may calculate the amplitude of vibration of the air from equation (4), Article 345, or $I = 2\pi^2 n^2 r^2 V d$, where, from equation (5), $I = 24 \times 10^{-9}$ erg/cm² sec.; $V = 344$ m/sec. at 20° C, $d = 1.2 \times 10^{-3}$ g/cm³ at 20° C, and $n = 512$ v.p.s. The resulting amplitude is 1.05×10^{-8} cm. This is of the order of magnitude of atomic diameters and gives an idea of the amazing sensitivity of the human ear.

There is also a superior limit to audibility, when the sound becomes so intense as to cause pain, and is no longer sound in the usual sense. This limit is associated with a vibration amplitude and excess pressure that for a tone of frequency 512 v.p.s. are about five million times their threshold values. The amplitude is about 2.5 mm, and the excess pressure is about 5 mm of mercury. Both the upper and lower limits of audibility vary greatly with the pitch, as we have seen, and differ with different individuals.

The following experimental values of sounds, measured in decibels above the average threshold value of sound, are quoted from a table by Dr. E. E. Free.†

	(Decibels)		(Decibels)
Painful sounds.	130-140	Vacuum cleaner.	70
Pneumatic riveter.	100-110	Average conversation.	65-75
Thunder.	80-110	Whispering.	25-30
Niagara Falls.	95	Heart beat.	10-15

358. Architectural acoustics. There are two principal causes of "bad acoustics" in an auditorium. One of these is the concentration of sound, due to one or more definite echoes, in certain regions and not in others. This depends upon the shape of the hall, and if it is complicated, as in a theater or gothic church, it is practically impossible to avoid such echoes deliberately, in the original design. But early in this century, Professor W. C. Sabine of Harvard University undertook the first really scientific study of acoustics, and succeeded in measuring the formation of echoes by means of a sectional model of the auditorium he wished to investigate. He used a device due to

† *Review of Scientific Instruments*, July, 1933.

Professor A. L. Foley of Indiana, by which one actually photographs the sound waves caused by a loud electric spark discharge at any desired time thereafter. The photograph is taken by the light of a second spark and is practically instantaneous. The wave creates an image on the plate, because wherever the air is compressed or rarefied, as the wave advances there is a distortion of the light from the spark source. In this way Sabine demonstrated the source of confusing echoes in the model, and thus the actual auditorium could be modified to eliminate them.

The second, and much more important cause of bad acoustics is *reverberation*, caused by so many echoes and re-echoes combining that they build up into what seems like a single continuous sound. If the walls of a room did not absorb some of the sound, a steady source would set up steadily increasing vibrations without limit. Actually there is such a limit, and a steady state is reached in a few seconds when the rate of absorption becomes equal to the rate of emission. Then if the source of sound is stopped, the reverberation dies down from its maximum value, following a logarithmic decay curve like the curve shown in Fig. 20, Article 331.

The reverberation time T required for a sound of definite energy density to die down to threshold value was investigated by Sabine, using a formula based on one discovered by W. S. Franklin in 1903. Sabine's standard source of sound developed an energy density (ergs per unit volume) one million times the threshold value, and the resulting equation was

$$T = 0.049 u / \Sigma(\alpha S), \quad (1)$$

where T is the reverberation time in seconds, u the volume of the room in cubic feet, S the areas of the various kinds of surface, and α is the absorption coefficient of each of these surfaces. This coefficient is defined as the ratio of the absorption of a given surface to that of an equal area of a perfect absorber such as an open window, which does not reflect, and may therefore be regarded as absorbing all the sound which reaches it. An open window thus has an absorption coefficient of unity.

The observed results agree remarkably well with the theory in all but very "dead" rooms where the total absorption is very large. The shape of the room has very little influence on the reverberation time, as Franklin had originally assumed, but T varies directly as the volume. This can be readily understood when we remember that absorption causing the original energy to decay occurs only when the sound waves meet the walls, ceiling, and so forth, and as the waves

are reflected back and forth between them, the larger the room is, the fewer are the reflections in a given time.

359. Calculations of reverberation time. It is often desirable to calculate T from the specifications of the auditorium before it is built. In this case we must know $\alpha_1, \alpha_2, \alpha_3, \dots \alpha_n$ for the various surfaces, as well as their areas $S_1, S_2, S_3, \dots S_n$. Then $\Sigma(\alpha S) = \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \dots \alpha_n S_n$, as explained in the preceding article.

The values of α for different substances depend upon the pitch of the sound, and these have been determined for a great variety of materials at different frequencies by Dr. P. E. Sabine of the Riverbank Laboratories. As an illustration, for smooth lime plaster on wood lath, α increases rather steadily from 0.024, with a tone of frequency of 128 v.p.s., to 0.037 at 1024 v.p.s. It then decreases to 0.019 an octave higher, but at 4096 v.p.s. it has increased again to 0.034.

Values of α determined by W. C. Sabine with constant pitch of average frequency are as follows:

Wood sheathing.....	0.061	Brick set in cement.....	0.025
Window glass.....	0.027	Audience, compact.....	0.96
Plaster on lath.....	0.034	Carpet (average).....	0.18

The absorption by a great variety of other surfaces has been measured, so that if the individual areas are known, it is possible to calculate the total average absorption, and consequently T , with some precision.

The ideal time of reverberation depends upon what the auditorium is to be used for. As a result of actual tests with musicians as judges, Sabine concluded that 1.1 second is the best value when listening to a piano in a rather small room. But F. R. Watson has shown that this may be increased in large halls. Watson also demonstrated that the best results are obtained when most of the absorption due to hangings or other absorbent material is near the audience, leaving the walls near the source of sound to create most of the allowable, and in fact desirable, reverberation.

If a hall is to be used for speaking, the consecutive syllables, if blended by reverberation, build up an increasing volume of sound by a succession of impulses until the steady state is reached, when the energy curve looks like the teeth on a broad saw, with very slight decrease of sound between syllables. This shows why they become blurred and indistinguishable. According to V. O. Knudsen, who

experimented in five empty high-school auditoriums, 2.75 seconds is the maximum reverberation time allowable in halls of this size, though it should be remarked that an audience, by increasing the total absorption, would have considerably reduced the reverberation time actually observed.

SUPPLEMENTARY READING

N. W. McLachlan, *Noise*, Oxford University Press, 1935.

P. E. Sabine, *Acoustics and Architecture*, McGraw-Hill, 1932.

F. R. Watson, *Sound* (Chapters 15, 16), Wiley, 1935.

—, *Acoustics of Buildings*, Wiley, 1930.

Steward and Lindsay, *Acoustics* (Chapters 9 and 11), Van Nostrand, 1930.

A. B. Wood, *A Textbook of Sound* (Section 4), Macmillan, 1932.

PROBLEMS

1. The faintest audible sound having a frequency of 512 v.p.s. has an intensity of about 24×10^{-16} watt/cm². How much louder is a sound whose intensity is 0.012 watt/cm²? *Ans.* 127 decibels.

2. The effective value of the excess pressure of a certain sound is 0.05 dyne/cm². Calculate its intensity, and its loudness compared to standard threshold excess pressure at 20° C, using equation (5) of Article 345. (Take $v_{20} = 344$ m/sec.) *Ans.* $I = 6.05 \times 10^{-12}$ watt/cm²; $b = 34$ decibels.

3. Calculate the reverberation time in a room 20' \times 30' \times 10' finished as follows: The walls are sheathed with wood halfway up, and plaster on lath covers the rest of the walls and ceiling. There are five open windows measuring 4' \times 3' in the upper half of the walls. The entire floor is carpeted. *Ans.* $1\frac{1}{4}$ sec.

CHAPTER 27

The Physical Basis of Music

360. Musical intervals. The ratio of the fundamental frequencies of two musical tones is called an **interval**, and it is customary in physics always to put the higher frequency in the numerator, so that intervals are never less than unity. If one note has a frequency $n = 72$ v.p.s., and the other $n' = 80$ v.p.s., the interval between them is $n'/n = 10/9$. In music, then, an interval *never* means a difference, such as $n' - n$. The simplest possible interval is $1/1$ or **unison**. The next is $2/1$, when $n' = 2n$, and this interval is called an **octave**. If we continue using integral intervals, we obtain a series of frequencies $n, 2n, 3n, 4n$, and so forth. The interval $4/1$ is a double octave, because it is an octave above $2n$, which in turn is the octave of n . The wave lengths of these tones vary inversely as n ; therefore if λ is the wave length of the lowest tone, $\lambda/2$ is the wave length of its octave, $\lambda/3$ of the next higher in the series, and so on. This series, $\lambda, \lambda/2, \lambda/3, \lambda/4, \dots \lambda/n$, is known as an harmonic series.

The simplest possible fractional interval greater than unity is $3/2$. If $n' = 3n/2$, then $\lambda' = 2\lambda/3$. This interval is known as a **fifth**. The next simplest is $4/3$, known as a **fourth**, with others following as listed below:

Interval	Name	Interval	Name	Interval	Name
1/1	unison	5/3	major sixth	10/9	minor tone
2/1	octave	5/4	major third	15/8	major seventh
3/1	twelfth	6/5	minor third	16/15	limma
4/1	double octave	8/5	minor sixth	25/24	diesis
3/2	fifth	9/5	minor seventh	81/80	comma
4/3	fourth	9/8	major tone		

These intervals, until we reach $8/5$, are the simplest possible prime fractions arranged in the order of decreasing simplicity. No ordinary musical intervals use prime numbers higher than 5; therefore such possible intervals as $7/5$ do not appear. They could be produced, but would not be pleasing to the ear. The reason for most of the names of the common intervals will be explained later.

361. The diatonic scale. Modern occidental music is based upon a gamut, or scale, of eight notes in a sequence of tones whose frequencies increase as we ascend the scale, and end with the eighth note having twice the frequency of the first, thus completing an *octave*, or a series of *eight* notes. The choice of these eight notes is far from arbitrary, although it seems to satisfy the ear in a manner not to be accounted for by purely mathematical reasoning. But at any rate, the choice involves ratios of the smallest integers, and therefore has a fundamental quality which seems somewhat inevitable. The series of notes are named *do, re, mi, fa, sol, la, ti, do*, pronounced as in Italian. The first interval between *re* and *do* is $9/8$, or a *major tone*. The second, between *mi* and *do* is $5/4$, or a *major third*, because *mi* is the third note of the scale. Between *fa*, the fourth note, and *do*, it is $4/3$, a perfect *fourth*. Between *sol*, the fifth note, and *do*, the interval of $3/2$ is a perfect *fifth*. Between *la* and *do* it is $5/3$, a *major sixth*, and between *ti* and *do* the interval of $15/8$ is a *major seventh*. Thus all the notes of the gamut, except the seventh, are expressed in terms of the simplest possible ratios.




The intervals between successive notes may be obtained from the above ratios as follows: Suppose we wish the interval between *mi* and *re*; then as *re* is a major tone above *do*, the interval $re/do = 9/8$. But *mi* is a *third* above *do*; therefore $mi/do = 5/4$. But if the second equation is divided by the first, the *dos* cancel, and we have $mi/re = 5/4 \times 8/9 = 10/9$, which is the interval sought, a minor tone. In this way the following table may be calculated, when *do* is taken as the *C* of ordinary musical notation, and the series is known as the **diatonic** (thorough-toned) **scale**: Intervals between each note and *do* are given, as well as the intervals between successive notes.

Name	<i>do</i>	<i>re</i>	<i>mi</i>	<i>fa</i>	<i>sol</i>	<i>la</i>	<i>ti</i>	<i>do</i>
Letter	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>A</i>	<i>B</i>	<i>c</i>
Interval above <i>do</i> .	1	$9/8$	$5/4$	$4/3$	$3/2$	$5/3$	$15/8$	2
Successive Intervals.	$9/8$ $10/9$ $16/15$ $9/8$ $10/9$ $9/8$ $16/15$							

If, for example, *do* is 24 vibrations per second, the various frequencies may be obtained by multiplying 24 by $9/8$, then by $5/4$, then by $4/3$, and so forth, or what is the same thing, each note may be found

from its predecessor by using one of the intervals in the lower row. This gives the following frequencies for the notes of this octave:

<i>do</i>	<i>re</i>	<i>mi</i>	<i>fa</i>	<i>sol</i>	<i>la</i>	<i>ti</i>	<i>do</i>
24	27	30	32	36	40	45	48

It is customary to denote the octave , which starts with *C* below the bass clef, by the capital letters of the alphabet. The next higher octave, , is indicated by the lower-case letters, *c, d, e*, and so forth, and the octave  is indicated by these letters primed, *c', d'*, and so forth, the next higher by double primes, and so on, while the octaves lower than *C* are denoted by subscripts.

362. The triads. It is also possible to build up the scale from certain fundamental groups of these notes known as **triads**. When played together they constitute **chords**, and are the basis of all *harmony* in music, as distinguished from pure *melody*, when there is only a sequence of tones. There are three *major* triads, and three *minor* triads, as well as *diminished* and *augmented* triads. The three major triads all have the ratios 4:5:6, and the minor triads, 10:12:15. An augmented triad is a major triad with the upper note sharpened. A diminished triad is a minor triad with the upper note flatted. The three of each set are known as the **tonic, dominant, and subdominant**.

In the key of *C*, they are as follows:

Major	$\left. \begin{array}{l} \text{Tonic } C : E : G \\ \text{Dominant } G : B : d \\ \text{Subdominant } F : A : c \end{array} \right\}$	4 : 5 : 6
Minor (descending scale)	$\left. \begin{array}{l} \text{Tonic } C : E^b : G \\ \text{Dominant } G : B^b : d \\ \text{Subdominant } F : A^b : c \end{array} \right\}$	10 : 12 : 15

The small letters show, as just explained, that the note is in the higher octave.

It will be seen that the dominant begins its triad where the tonic leaves off, and the subdominant works backward from *c*, with a result that this triple application of the ratios 4:5:6 gives us all the notes of the major scale.

The minor triad introduces three new notes, *E* flat (*E^b*), *A* flat (*A^b*), and *B* flat (*B^b*). These are each lower than the corresponding notes of the major scale by an interval of 25/24, called a **diesis**. The

frequency of the lower note is $24/25$ of the upper one. This gives rise to a new (minor) scale which has actually two different forms or modes. One, called the **harmonic mode**, uses E^b and A^b but not B^b , both ascending and descending. The other, or **melodic mode**, uses only E^b ascending, and all three flats descending. The successive intervals of the descending melodic minor in C are given below.

C	D	E^b	F	G	A^b	B^b	c
1	$9/8$	$6/5$	$4/3$	$3/2$	$8/5$	$9/5$	2
	$9/8$	$16/15$	$10/9$	$9/8$	$16/15$	$9/8$	$10/9$

The new intervals shown in the upper row, that is, $6/5$, $8/5$, and $9/5$, are the minor third, minor sixth, and minor seventh, respectively.

It will be noticed that there are only three numerically different intervals between successive notes in both major and minor scale. Two are nearly alike, $9/8$ and $10/9$, major and minor tones; and the third is the much smaller interval $16/15$, or **limma**. In the major scale, the third and seventh intervals are limmas, but in the minor scale they occur in a different order, according to the mode, giving a characteristic mournful value to music played in a minor key.

363. Standards of pitch. It is purely arbitrary what particular frequency we shall assign to the note known as middle C (c'), from which other tones above and below it are calculated. Physicists commonly assign it 256 v.p.s. But the same note has 264 v.p.s. in "orchestral pitch," as used by most musicians. This is based on an A (a') having 440 v.p.s., for it is the A that is given out by one instrument of the orchestra for all the others to tune to. There is still another and older pitch known as "concert pitch" whose A is 460 v.p.s., giving a C of 276 v.p.s., which is nearly a whole tone higher than the middle C of the physicist's scale.

Evidently there is no pitch belonging naturally to the note we choose to call C , for this note, played on an old-fashioned piano, would have almost the same pitch as D of the physicist's scale, or $C\#$ of orchestral pitch.

Having fixed C (or A) of a scale, all other notes up and down are also fixed, but the sequence of tones, *do*, *re*, *mi*, and so forth, may be applied from any starting point, and is thus wholly flexible. Then this "movable *do*" may mean a note of any frequency which is used as the beginning of the scale or gamut. There is, however, another system, dating from the eleventh century, which uses a "fixed *do*" so that *do* would always mean C , *re* would always mean D , and so on,

thus making these names practically synonymous with the alphabetical notation.

364. Consonance and dissonance. If two notes have nearly the same frequencies, the slow beats which are produced when they are sounded together are rather agreeable. In fact, this effect is made use of in the "vox celeste" stop of the organ, which causes two pipes of slightly different pitch to "speak" together when a key is struck, producing a pleasing pulsation. If, however, the beating becomes more rapid, a disagreeable sensation is produced by the rapid flutter of the beats, and even when they are too rapid for the ear to follow as separate impulses, there is still a displeasing roughness between the notes, owing to their *dissonance*. If the difference of the frequencies is then still further increased, the dissonance gradually disappears, and we approach *consonance*, which gives a pleasing effect when two notes are sounded simultaneously. The number of beats per second which is unpleasant depends upon the absolute pitch of the notes which produce them. Two low notes near *C* below the bass clef, when they produce beats at a rate of from 4 to 14 per second, are dissonant, while two notes whose frequencies are near high *C* just above the treble clef must beat from 40 to 100 times per second to sound unpleasantly together.

A pure tone (devoid of overtones) is always consonant with its harmonics. In this case the notion of beats is rather meaningless. A note and its twelfth, as 60 and 180 v.p.s., beat theoretically $180 - 60 = 120$ times per second, which is the octave of the lower note. But since octaves are consonant, this beating does not produce dissonance. By the same kind of reasoning it may be shown that even the higher harmonics are all consonant with their fundamental. However, the harmonics of a note are not necessarily consonant with each other; thus if the fundamental has 80 v.p.s., the twelfth harmonic has 960, and the thirteenth, 1040. The difference 80 is the frequency of the beats between them, and lies within the region of dissonance (40 to 100) at these frequencies just above the treble clef. Thus it is evident that a complex tone may have overtones which are dissonant with each other, and it has therefore a less pleasing sound than one whose overtones include only consonant harmonics.

The question of consonance or dissonance between notes is complicated by the fact that we have to consider the overtones present in complex tones. Then intervals otherwise consonant may become dissonant, owing to beats between the overtones of the notes concerned. In general, the overtones of the two notes forming a highly

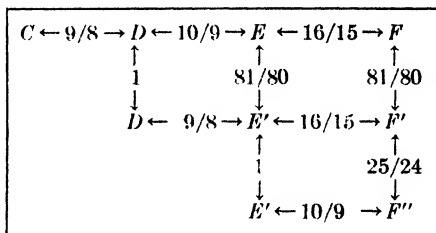
consonant interval such as the octave, are consonant with each other, but even in this case the higher overtones may be dissonant, though they are never strong enough to be observable. Thus the fifteenth harmonic of a note whose frequency is 60, has 900 v.p.s., and the eighth harmonic of its octave (120 v.p.s.) has 960. Their difference, giving 60 beats per second, is dissonant *at this pitch*.

In the case of intervals less than an octave, the lower and therefore stronger overtones may develop beats which render such combinations dissonant even when their fundamental frequencies would otherwise be consonant. Thus a note whose frequency is 72 and its major third, 90 v.p.s., may have as overtones the harmonics listed below:

Fundamental	2nd	3rd	4th	5th Harmonic
72	144	216	288	360
90	180	270	360	450

Here two of the harmonics are in unison; namely, the fifth of 72 and the fourth of 90. Others like 216 and 180 are harmless because their difference of 32 is too fast to be displeasing at the pitches they represent. But the fourth of 72 and the third of 90 together constitute the interval of a limma, $16/15$, and produce 18 beats per second, which is within their discordant region. Therefore if the two notes have these overtones strongly marked, they will not sound well together.

365. Transposition. In order to make it possible to play a diatonic scale beginning with any of the notes of the key of *C* which we have so far exclusively considered, new notes involving new intervals must be introduced. Suppose, for instance, we wish to raise the pitch of our gamut by a major tone, beginning with *D* as the keynote. Then we must have a new *E*, because the old *E* is only a minor tone above *D*,



Relation between Notes Needed in Transposition

whereas the first interval of a diatonic scale must be a major tone. If we denote this new note by *E'*, its relation to *D* is given by $E'/D = 9/8$. But $E/D = 10/9$; therefore, dividing the former relation by the latter, we have $E'/E = 81/80$, or a **comma**, an

interval perfectly recognizable by those having acute musical perception. This would raise the next note F on such a scale by a comma also, if F were still a minor tone above E' . But this is not enough, since $F/E = 16/15$, a limma, whereas we now require a minor tone, or $10/9$. Therefore the interval between F' and the required note, F'' , is obtained as before by dividing $F''/F' = 10/9$ by $F'/E' = 16/15$, giving $F''/F' = 25/24$, or a diesis. Thus, augmented by a semitone, F'' is called F sharp, denoted by $F\sharp$. The interval F''/F should really be called a *major diesis*, since it is the ratio between a major tone and a limma, that is, $9/8 : 16/15$, which equals $135/128$. Similarly F''/F' is a *minor diesis*, which is the ratio of a minor tone to a limma and equals $25/24$.

In the key of D , transposed as above, we have introduced two new notes; namely, E' and F'' , not counting the little-used F' , but if the process is continued, several more will be needed before the octave is completed. The most important of these is $c\sharp$, which makes the interval between the last two notes a limma instead of the major tone between c and d .

In the same way we might start the scale with E , and by the introduction of another group of new notes, complete a perfect scale. In some scales, such as the key of F , a note must be "diminished" instead of augmented. Such a note is said to be the *flat* of its original pitch, and is obtained by dividing its frequency by the ratio $25/24$ or $135/128$, according to whether a minor or major diesis is called for.

If absolute accuracy were necessary in transposition, both the major and minor diesis would have to be used in forming the new scales, and each note of the natural scale would then have two flats and two sharps. As a result of this refinement, the sharp of a note may be either lower or very slightly higher than the flat of a note a minor tone above it. But if a major tone is involved, the *calculated* sharp of the lower note is invariably lower than the flat of the note above it, in spite of an apparent tendency on the part of some musicians to regard it as of a slightly higher pitch.

366. The tempered scale. An instrument played with keys, like the piano or organ, would have to have its keyboard greatly extended with additional notes in order to make it possible to play the diatonic scale in any key. A minimum of 53 such notes would be necessary for each octave, which would make it extremely difficult both to construct and perform on.

This problem was attacked by the great German composer Johann Sebastian Bach (1685–1750), and solved for all practical purposes by

his invention of an instrument which he called "the well tempered clavichord." This device, by a system of compromises, reduces the 53 different notes required for transposition, to the thirteen black and white keys of the octave of the modern piano. This he effected by establishing only two sorts of intervals between adjacent notes, one of which is a compromise between the nearly equal major and minor tones, and the other between the somewhat more different limma and diesis. Thus $E\sharp$ is made identical with F , and $B\sharp$ with C . The sharp of a note is made identical with the flat of the one next higher, and the difference between a major and minor tone vanishes in a compromise interval called a **whole tone**.

This results in a sequence of the twelve intervals of the "chromatic scale," which are known as **half-tones** or **semitones**, such that the product of two successive semitones ($C\sharp/C$) \times ($D/C\sharp$) results in the whole tone D/C . The product of the twelve semitone intervals is the octave of the original note, so that if the interval is denoted by x , and if it is applied twelve times to a note of frequency n , the result must be $2n$; hence $x^{12}n = 2n$, or $x = \sqrt[12]{2} = 1.0595$. All the intervals, with the exception of the octave on such a scale, are different from their ideal values. The *fifth*, G/C , for instance, involves a sequence of the seven semitones between C , $C\sharp$, D , $D\sharp$, E , F , $F\sharp$, and G . Therefore this interval on the tempered scale is $2^{7/12} = 1.0595^7 = 1.498$ instead of $3/2$, or 1.500 on the diatonic scale.

Music played on an instrument like the piano or harp, or on those whose pitch is definitely fixed by frets, as in the guitar, or by holes like those of the flute, is slightly inferior to music played in "just intonation." The violin and other stringed instruments without frets, also the slide trombone, and of course the human voice, are not so limited, and can therefore use the diatonic scale when not accompanied (except by each other), thus producing a more pleasing effect.

SUPPLEMENTARY READING

- C. G. Hamilton, *Sound and its Relation to Music* (Chap. 7), Oliver Ditson, 1912.
 F. R. Watson, *Sound* (Chap. 13), Wiley, 1935.

PROBLEMS

1. What is the interval between a note and a fifth above its octave? What is the interval between the major and minor sixth above a given note?
Ans. 3; 25/24, or a diesis.

2. If C has a frequency of 256 v.p.s., calculate the frequency of F^\sharp on the diatonic scale. *Ans.* 355.6 v.p.s.
3. Calculate the interval between F^\sharp and B^\flat on the diatonic scale. *Ans.* 1.296.
4. By what interval must F be augmented if a diatonic scale is to begin with E ? *Ans.* 135/128.
5. Calculate the interval between C^\sharp and F on the tempered scale. *Ans.* 1.26.
6. Calculate the interval between F^\sharp of the tempered scale and the same note of the diatonic scale. *Ans.* 1.018.

CHAPTER 28

The Production of Tones—Vibrating Solids

367. The vibration of strings. We have already seen that when two transverse vibrations of the same wave length and velocity, and traveling in opposite directions, meet each other, they combine in the production of standing waves. This occurs in a string stretched between two supports. If the string is set vibrating, the waves set up are reflected at the ends, and the reflected and original waves interfere, forming a node at each end and one or more loops in between.

In Fig. 35, a vibrating string is shown greatly exaggerated, in four phases a quarter of a period apart. Evidently *a*, *c*, *e*, and *g* mark nodes, while *b*, *d*, and *f* are antinodes. If only *a* and *g* were nodes with a single loop between them, the wave length would be twice the length of the string, which would then be vibrating with its lowest possible frequency, or in its fundamental mode. If there were two loops, the wave length would equal the length of the string, and the frequency would be twice that of the fundamental. In the case shown, there are three loops, and the wave length is two thirds the length of the string, then vibrating with three times the frequency of the fundamental, which is the third harmonic.

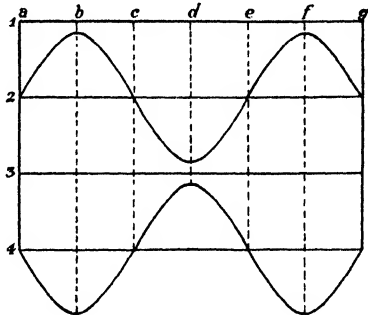


Fig. 35.

368. The velocity of waves in a stretched string. In order to calculate the frequency from a given wave length, it is necessary to know the velocity v with which the wave advances. This depends upon the tension in the string, and on its mass per unit length, as proved below.

Let us suppose that the top of an advancing crest is the arc ab of a circle of radius r whose instantaneous center is at O . This is not exactly the case for sine waves, but it is an experimental fact that the form of the wave does not affect v in either longitudinal or transverse

waves, so we may make the assumption without affecting the result. The particle p of the string in Fig. 36 is at the top of such a crest and is making vertical harmonic vibrations while the crest is moving to the left with a velocity v .

Let the mass of each centimeter length of the string be μ , and let the tension be T dynes. Let ab be a very small length l with

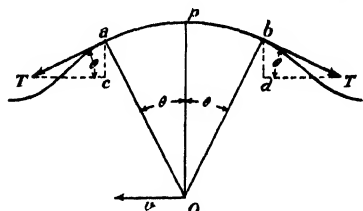


Fig. 36.

p at the center; then the mass between a and b is μl . Now it makes no difference in the calculation whether we regard the crest as running along the string, or the string running over the crest, so that we may treat this elementary mass μl as being acted on by a centripetal force directed toward O

and having a centrifugal reaction given by $F = \mu v^2/r$, where v^2/r is the acceleration toward the center.

The centripetal force which holds p in a circular path is supplied by the vertical component of the tension of the string acting at a and b . This tension acts tangentially to the curve, and its vertical components are ac and bd . The angles θ shown in the diagram are all equal because their sides are mutually perpendicular; therefore $ac = bd = T \sin \theta$. But $\theta = ap/r = bp/r = \sin \theta$ very nearly, because ab has been assumed very small compared to r . Then the total downward force due to the tension is

$$2T \sin \theta = \frac{2Tbp}{r} = \frac{Tl}{r}.$$

Setting this equal to the centrifugal reaction derived above, we obtain

$$\frac{\mu v^2}{r} = \frac{Tl}{r},$$

$$v^2 = \frac{T}{\mu},$$

and

$$v = \sqrt{\frac{T}{\mu}},$$

which is the equation sought, giving v in terms of centimeters per second, if the tension is expressed in dynes, and the mass in grams per centimeter of length.

369. Frequency of vibration of stretched strings. We may now calculate n from the velocity equation derived above provided the

tension and mass per unit length are known. When the string vibrates in its fundamental mode, its length, as we have seen, is equal to half the wave length, or $l = \lambda/2$, where l now means the total length. The next possible cases are $l = \lambda$ and $l = 3\lambda/2$, as we have also seen. When $l = 4\lambda/2$, there are four segments, and so on. In general, $l = k\lambda/2$, where k is any integer. Therefore $\lambda = 2l/k$. But v always equals $n\lambda$; hence $v = 2ln/k$. Equating this with the value for v obtained above, and solving for n , we obtain

$$n = \frac{k}{2l} \sqrt{\frac{T}{\mu}}, \quad (1)$$

Equation (1) gives the frequency as a function of the tension, length, mass of the string per unit length, and an integer k . This integer is unity when the string is vibrating in its fundamental mode (the usual case with stringed instruments), and 2, 3, 4, and so forth, when it is giving its second, third, and fourth harmonics.

These higher harmonics may actually be obtained by placing the finger lightly on a point of the string where a node is necessary for the number of segments required. Then the string is made to vibrate at or near an antinode, by bowing it there as on the violin. This "harmonic bowing" may produce higher notes than than would otherwise be possible, and is frequently used by violinists.

370. Melde's experiment. The principles involved in the equation (1) of the preceding article are illustrated in an ingenious device

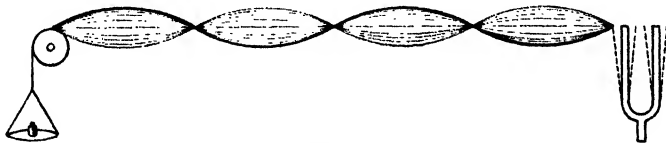


Fig. 37.

designed by Melde. This consists in imposing a constant period of vibration upon a string by means of an electrically driven tuning fork, but with the tension made variable by passing the end of the string over a light pulley and hanging weights from the end, as shown in Fig. 37. Then $T = mg$, where m is the mass, hung from the string, that may be varied at will. Equation (1) of the last article now reads

$$n = \frac{k}{2l} \sqrt{\frac{mg}{\mu}},$$

where l , g , and μ are constants. But the frequency of the string's vibration, which is half that of the fork when attached as illustrated,

is also a constant. Thus the only variables are k and m , and these are obviously so related that $k \propto 1/\sqrt{m}$, or $m \propto 1/k^2$. Therefore, since k can have only integral values, m is limited to values that correspond. These are clearly multiples of 1, $1/4$, $1/9$, $1/16$, and so forth. Thus if the total mass producing tension is 144 grams when the vibration is in its fundamental mode (one segment), 36 grams would give two segments, 16 would give three segments, 9 would give four, and so on.

Since the production of the segments depends upon both l and m , we may always obtain them with a given weight by varying the length of the string by a forward or backward motion of the pulley. When the correct position is reached, the segments appear as hazy spindles shown in the diagram, but the instantaneous position of the string is a continuous curve like the heavy line, where it is shown in maximum displacement.

371. Vibrating rods. Rods or bars of some sufficiently elastic solid may vibrate either transversely, longitudinally, or perform torsional vibrations. When they vibrate transversely, they may be clamped at one end while the other end vibrates. In this case several modes of vibration are possible, as with strings, but now only one end can be a node, because the free end must be in vibration. A second, third, or fourth node may occur elsewhere, as indicated in Fig. 38.

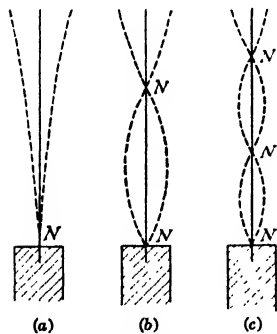


Fig. 38.

These modes of vibration do not produce the regular odd harmonics of the fundamental, as they would with a vibrating string having the same nodes and loops.

In the case of vibrating bars, the various frequencies depend upon the length, elasticity, and density of the bar. When clamped at one end, the frequencies of the first three overtones, referred to the fundamental as unity, are 6.267, 17.55, and 34.39, instead of 3, 5, and 7, as we might be led to expect from the apparent resemblance to vibrating strings, as shown in Fig. 38.

A thin strip of metal, clamped at one end and set vibrating by a blast of compressed air or steam, constitutes the "reed" of some kinds of foghorns or steam trumpets. These belong to the same class of instruments as the clarinet or saxophone, though in the latter group the reed is really a *reed*.

Another type of transverse vibration is produced when rods or bars are supported in a horizontal position at two nodal points like the

wooden bars of the xylophone, or the metal bars of the glockenspiel. Then they vibrate as shown in Fig. 39 (a). Fig. 39 (b) shows the rod producing its first overtone. If it were a string, this would be the octave of the vibration frequency of case (a), for the portion between the supports contains half a wave in (a) and a whole wave in (b). This is not the case, however, for the ratio of these frequencies is somewhat higher than an octave.

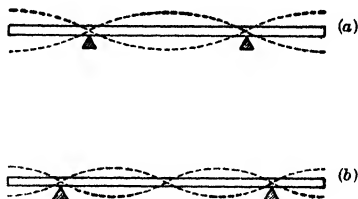


Fig. 39.

The tuning fork is really an application of a vibrating rod, supported at two points brought close together. This may be understood by imagining the rod in Fig. 39 being gradually bent into the form of a fork, while its two supports are brought closer and closer together, as shown in Fig. 40. Thus the shank of the fork *S* really represents the portion of the straight rod between the supports which still continues to vibrate up and down. These vertical vibrations are often used to set in vibration the sounding board of a resonance chamber on which the fork is mounted. If the enclosed air column is of the proper

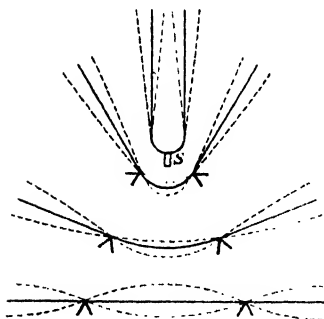


Fig. 40.

length, the sound is greatly increased.

Both stretched strings and rods may be made to vibrate longitudinally, exactly as if a sound wave were passing over them and interfering with a reflected wave moving in the opposite direction, thus setting up standing longitudinal vibrations with nodes at the fixed points. In fact, these vibrations are really stationary sound waves set up in the medium considered. Thus a rod clamped at one end, as in Fig. 41 (a), may be made to vibrate in this way by rubbing it

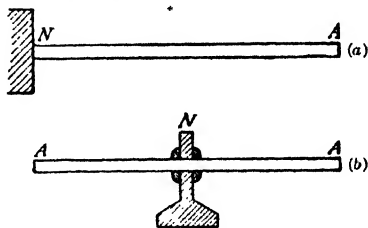


Fig. 41.

endwise with a piece of leather sprinkled with rosin dust. The irregular friction set up tends to extend and then release it, and the

resulting standing wave has a node at the clamp, and an antinode at the free end. The wave length is therefore four times the length l , and the frequency is obtained from $V = \lambda n = 4ln$, where V is the velocity of a longitudinal wave (that is, sound) in the material of the rod.

If clamped at the center as in (b), $\lambda = 2l$, so that the two antinodes are only half a wave length apart, and are therefore in opposite phase. Then the two ends move either toward or away from each other simultaneously, and the bar alternately contracts and expands in length. This arrangement is used to determine the velocity of sound in metals by measuring λ and n with an apparatus known as Kundt's tube, to be described later.

A rod clamped at one end may also be made to twist and untwist, thus setting up torsional vibrations, with a node at the clamp, and an antinode at the free end. It is even possible to produce additional nodes which result in the higher frequencies of the overtones. The theory of such vibrations involves the torsional rigidity of the rod, and is therefore not so simple as that of the other two types.

372. Vibrating plates. A thin metal plate held rigidly at some fixed point may perform transverse vibrations of one or more definite frequencies. In general, this fixed point is either the center of area or the entire periphery. Square or triangular plates clamped at their centers perform very complicated vibrations in which definitely limited and adjacent areas are in opposite phase with nodal lines between them.

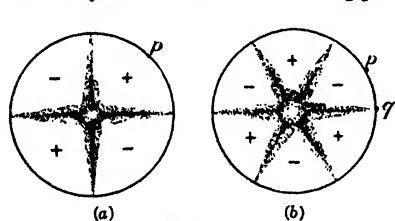


Fig. 42.

From a practical point of view, a circular disc is the most interesting form. If clamped at its center and bowed across the edge at a point p (Fig. 42 (a)), it vibrates in four sectors with nodal lines between them. The adjacent sectors are in opposite

phase as indicated by the $+$ and $-$ signs, and are of course moving perpendicularly to the plane of the disc. Such vibrations may be conveniently studied by sprinkling sand over the plate. When it is set in vibration, the sand is violently agitated, and thrown away from the vibrating areas into the nodal regions, where it collects and forms a characteristic pattern.

The sand-pattern method of studying the vibrations of plates was devised by Chladni, a German physicist of the 18th century. The patterns produced in this way are called Chladni's figures. Some of

them are very complicated and beautiful. The case (*a*) in Fig. 42 is the fundamental mode of vibration of a circular disc with immovable center. It may, however, be made to produce overtones by holding it rigid at some point *q* (shown in (*b*)). This point, in conjunction with the center, determines a nodal line. The edge is then bowed at a point *p* whose distance from *q* divides the circumference into an even number of equal parts. In the case shown, *pq* is one twelfth of the circumference, and six vibrating sectors are produced with radial nodes between them. Other possible patterns, obtained by W. S. Franklin, are sketched in Fig. 43.

If the disc is held rigidly around its edge, like the diaphragms of the telephone or phonograph, the fundamental mode is obviously one

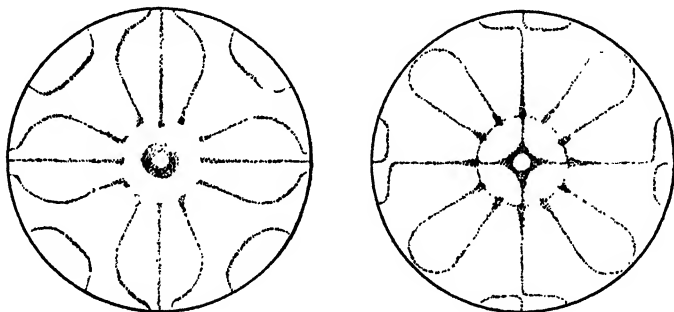


Fig. 43.

in which the whole surface becomes alternately convex or concave upward, with the maximum displacement at the center. Such diaphragms have a definite period of free vibration, and if this period coincides with that of the impulses causing it, the resulting amplitude and sound emitted are greatly increased. Resonance of this sort in telephone and phonograph diaphragms must be avoided within the range of frequencies of the voice or musical instruments, otherwise a strong distortion of the sound would be produced at or near the critical frequency. Harmonics of such discs are produced when the vibrating areas are zones with concentric circles as nodes between them, or are sectors with radial nodes, as in Fig. 42, or are both combined.

Vibrating membranes, like drum heads, behave in a similar but more complicated manner, when the natural period of their fundamental tone coincides with the resounding air chamber behind them. This is true of the tympani (kettle drums) of the orchestra, whose fundamental frequency becomes strongly predominant through resonance, and a tone of very definite pitch is produced. The pitch

of ordinary drums may be varied by varying the tension of the membrane, or by loading it with a heavy paste over its center, the latter being a method commonly practiced in some oriental countries.

Bells may be regarded as vibrating circular plates deformed by curving them over a more or less conical surface. The top of the bell, or center of the plate, is always a node, and the transverse vibrations of the rim are at right angles to the bell's axis. The simplest or fundamental mode of vibration corresponds to Fig. 42 (a), and results in deforming the rim into an ellipse, with four nodes where it intersects the circle of the undeformed rim, as in Fig. 44. If the bell is

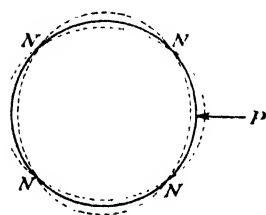


Fig. 44.

struck at P , the major axis is at first vertical, but the return vibration carries it past the zero position to form another ellipse having a horizontal major axis. It then oscillates with a definite frequency between these two positions. The production of overtones, as with flat discs, is very complicated, and the wave lengths do not form an harmonic series.

This results in dissonance between some of them, and the art of the bell founder consists in eliminating those which are undesirable. But the theory of such vibrations is, generally speaking, too complicated for analytical solution, and only gradually acquired experience, largely through a method of trial and error, has evolved the highly technical art of campanology.

373. Measurement of pitch. The determination of the pitch of a vibrating body really means comparing it with another vibration of known frequency. We may, for instance, determine the pitch of a tuning fork by means of a stroboscope. This is an instrument in which a disc with a ring of holes, as in the siren, is set rotating, and the fork is either viewed through the holes or is illuminated by an intermittent beam of light passing through them. If one hole replaces the next one in front of the eye (or beam of light) in the time required for a complete vibration of the fork, it will appear to be at rest, because the instantaneous pictures thus obtained are all alike. Then if the number of holes in the circle is known, and the number of revolutions the disc executes in a second, we can compute the time of replacement and so find T and n . If the speed is twice as fast, then every other hole will give us the same picture, but the intermediate ones show the fork in opposite phase. In this case the fork still seems at rest, but looks like a "double exposure" of two opposite positions, unless they happen to be those of mid-swing. Therefore, to deter-

mine n , the disc should be speeded up gradually until the fork first appears to be at rest, when the calculation indicated above is valid.

374. Compounding vibrations normal to each other. We have so far discussed only those cases of compound vibrations in which the components differ in frequency, phase, or amplitude. They may, however, vibrate in different directions as well. The most important case occurs when two vibrations at right angles to each other are compounded into a single vibration, a condition common in many mechanisms. When the two vibrations have the same amplitude and frequency, the result of their displacements traces out a circle if the two vibrations are in phase quadrature, and a straight diagonal line if they are in the same phase. The first case is readily understood by referring to Fig. 69, Article 91, where the particle P' is considered as the source of two harmonic motions in phase quadrature along the two axes. Therefore the vector sum of the x and y displacements is the radius vector r , whose terminus sweeps out a circular path regarded as the *result* of the two combined vibrations instead of the *cause*. This may also be proved by taking the vector sum of $x = r \cos \omega t$ and $y = r \sin \omega t$, which is given by

$$\sqrt{x^2 + y^2} = \sqrt{r^2 \cos^2 \omega t + r^2 \sin^2 \omega t} = r.$$

The second case is almost self-evident, for if the two particles P start their excursions simultaneously from the origin and in the same or opposite phase, their vector sum is a variable length equal to $\sqrt{x^2 + y^2}$, making an angle of 45° with the axes, and having a maximum value, when $x = y = r$, equal to $r\sqrt{2}$.

If the two vibrations are neither in quadrature nor in the same or opposite phase, the result is an ellipse, with the circle or straight line

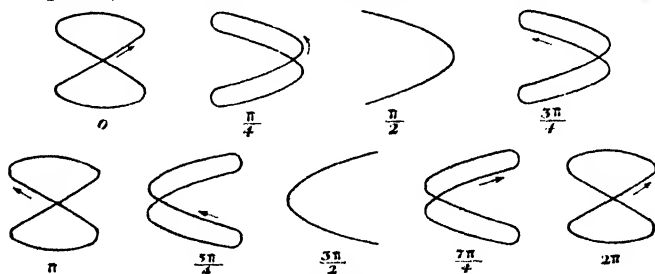


Fig. 45.

as extreme cases. Also, if the amplitudes are not the same, the circle becomes an ellipse, and the straight line is inclined at some angle other than 45° .

When two harmonic vibrations of different frequencies, but at right angles to each other, are combined, the resultant is a more or less complicated closed curve. These curves are known as *Lissajous' figures*. An interval of an octave gives a figure eight when the phase

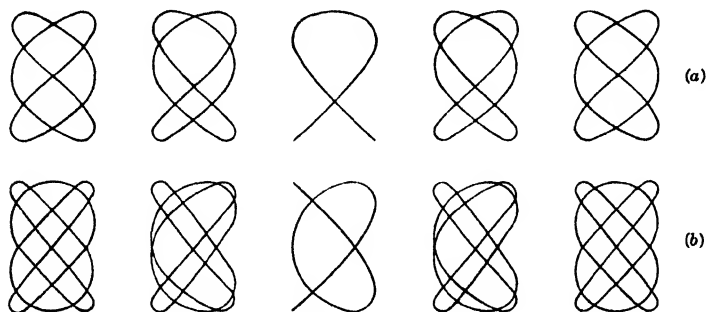


Fig. 46.

difference is either 0 or π . If it is $\pi/2$ or $3\pi/2$, the curve is a parabola, and for other phase differences the figures approximate one or the other extreme. These are shown in Fig. 45, where the phases differ progressively by $\pi/4$, and the arrows represent the direction of the resultant motion. An interval of $3/2$ gives rise to the figures of Fig. 46 (a), shown in successive quarter periods, and (b) shows the same series for an interval of $4/3$. The figures grow more complicated as the ratio of frequencies employs higher digits, and if musical tones are produced, these become less and less consonant, though the general character of the figures is always the same.

375. The piezo-electric oscillator. This remarkable source of sound vibrations is the result of a discovery made in 1880 by the French physicists, J. and P. Curie. They found that the opposite faces of

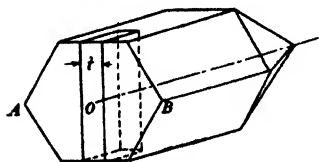


Fig. 47.

slabs, suitably cut from certain crystals, become oppositely electrified if the slab is stretched or compressed, and conversely, that when the faces are oppositely electrified, the slab changes in thickness and other dimensions.

Quartz crystals are most commonly used for this purpose. They have a more or less hexagonal section when cut across at right angles to the "optic axis," O, as shown in Fig. 47. If a slab of thickness t , having its faces normal to an axis such as AB, is cut from the crystal, the piezo-electric effect is a maximum, for

then a given pressure on the faces produces the maximum charge. The prefix *piezo* is derived from a Greek verb meaning "to press."

If a slab cut as described above is mounted between two metal plates, and if these are given a charge, the thickness t becomes either greater or less according to which face is positive. If an alternating voltage is applied, the crystal oscillates with the same frequency as the charging current. Now a quartz plate cut as specified has three natural vibration frequencies, the highest being associated with a vibration at right angles to the field, and if this frequency is applied by an oscillating electric field created in much the same manner as in radio transmission, the vibrations of the quartz build up to a much greater intensity than when they are "forced." This resonance frequency may be calculated from an empirical formula devised by A. Hund; namely,

$$n = 2.87 \times 10^5/t,$$

where t is the thickness of the slab in centimeters. Thus if the slab is 2.87 cm thick, its natural transverse vibration frequency is 100,000 v.p.s.

The resonance frequency of such oscillators is very sharply defined by a peak in the current curve due to interaction between the field and the vibration of the quartz plate. Professor W. G. Cady of Wesleyan University, who has exhaustively investigated the properties of these oscillators, found that the current falls to half its resonance value if the frequency varies 0.05 per cent in a 90,000 cycle oscillator. This property makes the device extremely valuable as a control of the frequency of radio broadcasting stations.

376. Supersonic vibrations. The vibrations created by the piezo-electric oscillator are usually of such high frequency that the longitudinal waves they send out in air or water are far above the audible range. This fact, and other valuable properties of short-wave sound, have led to much experimentation in their use for signaling and in depth-finding at sea.

In 1917, Professor Langevin of the Collège de France devised a supersonic oscillator, having a frequency of 50,000 cycles, which could be used under water both for transmitting and receiving signals. This device, as used on a ship, is essentially as shown in Fig. 48. A protecting box, D , projects from the hull of the vessel below the water line. In this box are a quartz mosaic slab, Q , and the metal plates A and B . The plate B is supported by a flexible rubber washer i , which insulates it but leaves it free to move, while plate A is also insulated but is held rigid. The source of electric oscillations,

E, is connected to the plate *A* by the insulated wire *c*, and to the other plate *B* by a wire *d* that projects into the water. As sea water is a good conductor, the circuit is thus complete, as indicated by the dotted line.

When *E* is properly tuned, *B* is set into rapid vibration and sends out very short waves into the water. At 14°5 C, the velocity of sound in sea water is 1435 m/sec., so with the frequency ordinarily used of around 35,000 cycles, the wave length is only 4.1 cm. Such very

short waves give sharp echoes from the bottom, and by measuring the time it takes for the "sound" to return to the ship, the depth may be calculated.

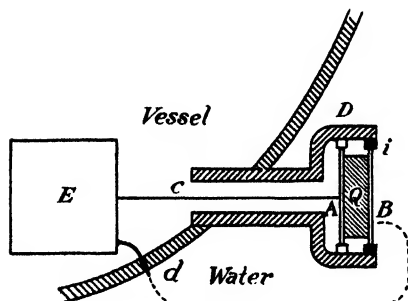


Fig. 48.

The quartz oscillator may also be used for submarine signaling between vessels many miles apart. The receiver, similar in construction to the transmitter, must be very closely tuned in order to respond. An

enemy vessel would thus have considerable difficulty in intercepting the signals. The high-frequency currents created in the receiving oscillator are converted into audio-frequency vibrations by the same devices as are used in radio reception.

Inaudible signaling through the air is also possible. The sound beam may be concentrated and directed by a parabolic reflector, and received in the same way. The energy contained in such a beam is very intense and may be picked up at much greater distances than would be possible with ordinary sound. It is, however, rapidly absorbed by the air under certain conditions, especially in the presence of an unusual amount of carbon dioxide. Then the range of signaling is greatly reduced.

377. The absorption of high-frequency sound in air. This phenomenon was practically discovered, and has recently been thoroughly investigated, by V. O. Knudsen of the University of California in Los Angeles. He finds that short waves may be absorbed far more than had been suspected from observations with usual wave lengths, and that the amount of absorption depends upon the temperature, relative humidity, and nature of the gas carrying the sound. The absorption, which rises steadily with the temperature, is far greater on a hot day than in very cold weather. It rises with the relative humidity up to a maximum somewhere between 10 per cent and

20 per cent, and then falls, so that at 92 per cent relative humidity the transmission of sound is as good as with perfectly dry air. Finally, Knudsen finds that this effect of moisture depends upon the oxygen in the air, rather than the nitrogen, while carbon dioxide, with suitable humidity, absorbs high-frequency sound even more powerfully than oxygen.

Knudsen's discoveries account in a great measure for certain well-known effects in nature. We can now better understand why sounds travel so much farther in the arctic than elsewhere, why we hear voices across miles of still water in cool, damp weather, and why the air over a hot, dry desert (humidity at say 15 per cent) transmits sounds so poorly. However, these illustrations refer to average frequencies, and are therefore not the best examples of a phenomenon that is particularly concerned with frequencies above 10,000 v.p.s. Such frequencies are still audible. They appear in consonant sounds like *s*, *th*, and *f*, and in the overtones which give timbre to musical tones. Therefore, in a hall where many persons are exhaling CO_2 , and where the air is moist and warm, much of the beauty of a master violinist's tone is lost, and a speaker may be hard to follow even if his diction is good and the hall acoustically perfect. The low-frequency harmonics reach the listener without serious absorption, but the higher harmonics of the vibrating string, and many spoken consonants are largely absorbed before reaching him.

SUPPLEMENTARY READING

F. R. Watson, *Sound* (Chapters 11, 12), Wiley, 1935.

J. W. Capstick, *Sound* (Chap. 11), Cambridge University Press, 1927.

E. G. Richardson, *Sound* (Chapters 3, 6), Arnold, London, 1927.

A. B. Wood, *A Textbook of Sound* (Section 2), Macmillan, 1932.

PROBLEMS

1. A wire weighing 48 g is stretched between two clamps 60 cm apart. The tension is 2×10^7 dynes. What is the pitch of the note produced when the wire is struck at its center? *Ans.* $41\frac{2}{3}$ v.p.s.

2. What tension is needed in the wire of Problem 1 if it is to give a note having a pitch of 60 v.p.s.? *Ans.* 4.1472×10^7 dynes.

3. In Melde's experiment, as shown in Fig. 37, a mass of 140 g results in four segments. What should be the mass in the pan to give five segments? *Ans.* 89.6 g.

* 4. An iron wire whose density is 7.7 g/cm^3 is stretched between clamps 90 cm apart. It is stroked *longitudinally* at its center and emits a note whose pitch is 2532 v.p.s. What is the elastic modulus of the wire? (NOTE: The waves set up in the wire are those of sound, as explained in Article 371.) *Ans.* 16×10^{11} dynes/cm².

CHAPTER 29

Production of Tones (*Continued*) Vibrating Gases

378. Vibration of an air column. Instead of vibrating strings, rods, and other solid bodies as a source of sound, standing waves may be produced in an enclosed column of air or other gas, which thus becomes the source of longitudinal sound waves. These waves are caused by interference between advancing and reflected waves, and their length is fixed by the dimensions of the column of vibrating air.

As in the case of solid bodies, something must start these vibrations, and a succession of periodic impulses is necessary to maintain them, just as the moving bow maintains the vibrations of the violin string. The actual source of the vibrations of an air column may be a vibrating air jet, or a vibrating solid such as a reed or stretched membrane.

379. Vibrating jets. If a blast of air is directed against a rigid wedge-shaped tongue of wood or metal, it tends to be set into rapid vibration back and forth across the edge of the tongue. This is illustrated in Fig. 49. The air blast comes through the chamber *C*, and is then forced through a narrow slit *S* toward the edge of *W* placed directly above it, which tends to divide it into two equal parts. But this exact equality is very unlikely to happen. If the portion *A* is a little the stronger, the air on that side will be more compressed than on the other, so that as the blast continues, it will become stronger on the *B* side, where

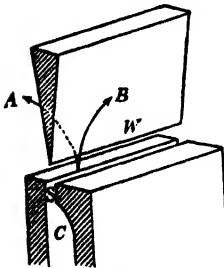


Fig. 49.

the pressure was lower. But this raises the pressure there and lowers it at *A* by the aspirator effect, so that the blast shifts back again to the first arrangement, but now with a more unequal division than at first. Thus after a few preliminary pulsations, it vibrates, almost as a whole, back and forth across the edge, as indicated by the curved arrows.

This vibrating jet has no very definite frequency and is extremely rapid. It sounds like the wind whistling around the corner of a house, a familiar sound produced in a similar manner. This rather weak and fluctuating vibration of the jet is used to start and maintain the sonorous tones of great organ pipes, as will shortly be explained.

380. Vibrating reeds. If the wedge described above were a thin flexible reed, it would be set in vibration by the jet of air, and so might serve as the source of vibrations of an air column, without the aid of the wigwag motion of the jet which drives it. Of course the blast which keeps a reed vibrating must necessarily vibrate also, but now the reed itself gathers energy and soon sets up much more vigorous "sympathetic" vibrations than those of the air blast. Their frequency, in the case of a free reed, depends upon the reed's rigidity and dimensions.

When the reed is vibrating, it sends out a train of longitudinal waves, and if they advance along an enclosed air column and are subject to reflection, standing waves are set up whose vibrations are the final source of sound in such a system. Thus an air blast (the original cause) sets up vibrations in a reed, which in turn sets up vibrations in an air chamber, and these launch a train of waves toward the observer's ear.

381. Resonance. We have already discussed resonance in showing how free vibrations are maintained by a correctly timed impulse. But it should be said that when so maintained, these vibrations are not entirely free, and there may be a very slight alteration in frequency, as with an electrically driven tuning fork whose natural frequency is slightly altered by the driving mechanism. However, synchronism between two vibrations, one of which sustains the other, is the necessary condition of resonance. When this is established, the body whose vibrations are thus maintained is said to *resound* to the other. In this way one tuning fork may set another vibrating by resonance, when they have exactly the same pitch or some integral multiple thereof. This is possible, though difficult to obtain, with only air as the intervening medium. But if the forks are mounted on the same sounding board, it is easily achieved provided the tuning is accurate.

If a note is sung, or produced in any way, near a piano having its dampers removed from the strings by holding down the heavy pedal, that string which has the same frequency as the tone responds in a striking manner, while other strings representing harmonics of the note will be found to "speak" also.

Air columns, also, resonate to an impressed vibration if their natural period is the same as that of the source. This property is made use of in mounting tuning forks on boxes open at one end. The partly enclosed air column has a definite natural period of vibration that amplifies the sound by resonance, if its period is the same as that of the fork.

The principal conditions which determine the resonance of cylindrical air columns may be illustrated by setting a tuning fork in vibration above a vertical tube open at the upper end, but closed below by a water column whose level may be altered at will. This is shown in Fig. 50.

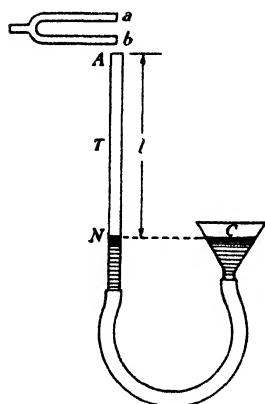


Fig. 50.

By raising or lowering the container *C* connected to the glass cylinder *T* by a rubber tube filled with water, the length *l* is varied. When the fork vibrates, there is a maximum of disturbance created at the mouth of the tube, and a minimum at *N*, where the surface of the water prevents longitudinal motion of the air molecules. Therefore, to establish resonance, the air column must have such a length that a standing wave is produced by reflection, with a node at *N* and an antinode at *A*. The minimum distance that realizes this condition is obviously a quarter wave length, and the fact that the fork

maintains the vibrations is shown as follows: A condensation sent out by the downward swing of the lower prong *b* is reflected at *N* as a condensation, and arrives at *A* after going a distance $2l$. This condensation is next reflected as a rarefaction, because sound travels faster in open air than in a tube, as was demonstrated by Regnault. So the air outside the tube is equivalent to a medium in which the wave travels more freely and which is virtually less dense than the air within the tube. If the prong *b* is now moving up, it helps in the formation of the reflected rarefaction, which is then returned to *N* to be again reflected, this time as a rarefaction that must reach *A* as the fork's lower prong starts its next downward swing. This rarefaction is reflected as a condensation and starts back toward *N*, amplified by the fork as before. Thus if the fork executes a complete cycle while the longitudinal disturbances it creates travel the length of the air columns four times, the two vibrations are in resonance. Therefore $4l_1 = VT = \lambda$, and $l_1 = \lambda/4$ as already stated.

If the water level is lowered until $l_2 = \frac{3}{4}\lambda$, resonance is again established, for this is the next shortest length of the air column which permits of a node at N and an antinode at A . The next possible case is $l_3 = \frac{5}{4}\lambda$. Then the wave length of the sound may be obtained from $\lambda = \frac{4}{1}l_1 = \frac{4}{3}l_2 = \frac{4}{5}l_3 = \frac{4}{7}l_4$ and so forth, and its frequency is obtained from $n = V/\lambda$, giving $n = kV/4l$, where k is any odd number. These various cases are shown in Fig. 51 (a), where the length of the arrows represents the direction and amplitude of vibration at various points along the tube, and Fig. 51 (b) shows the same thing by means of curves that represent the standing waves as if they were transverse instead of longitudinal.

Although the amplitude of vibration at a node is zero, the changes in pressure there are a maximum. This is because the molecules on either side of the node move in opposite directions at the same time, and so alternately crowd in toward it, or swing away from it. The reverse is true at an antinode, where there are no changes in pressure

because the molecules near it swing together in the same direction at substantially the same velocity without forming an appreciable condensation or rarefaction.

382. Organ pipes. A true organ pipe is a tube in which the air column is set vibrating by a vibrating air jet, and it may be either open or closed at the far end. It is essentially a large whistle, whether open like the usual whistle, or closed like those fitted with a piston to vary the pitch. The production of the tone by forcing an air blast against an edge has already been explained. The length of the air column determines the pitch by forcing the jet to vibrate in synchronism with the natural period of that column or with one of its harmonics. In the case of closed pipes, shown in Fig. 52, the frequency given by $n = kV/4l$ may be the fundamental tone when $k = 1$, or an overtone when $k = 3, 5, 7$, and so forth. These values give frequencies in the ratio of $1:3:5:7$ and so forth, a series of odd harmonics. Such overtones, always present to some extent when the fundamental is sounded, may be made predominant one after the other by increas-

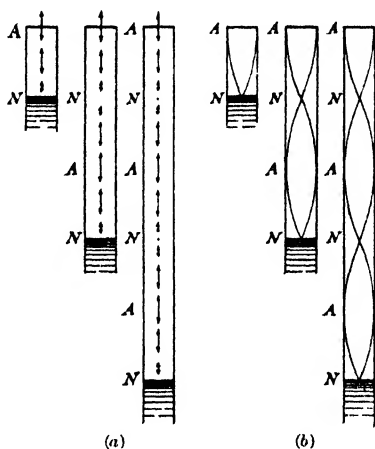


Fig. 51.

ing the pressure of the blast. Forcing the standing wave to break up into several segments is most easily accomplished with pipes whose cross section is small compared to their length.

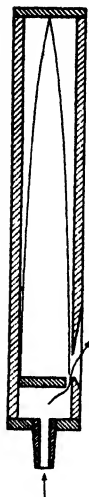


Fig. 52.

Open pipes give both even and odd harmonics. In this case the reflection at the open end, farthest from the vibrating jet, is such that a condensation is reflected back into the tube as a rarefaction, and vice versa. Consequently the original wave, after one reflection, returns to the starting point ready to be reflected in the same phase as at first. Thus it travels the length of the tube only *twice*, while the vibrating source has performed a complete cycle. Therefore $\lambda = 2l$ instead of $4l$, as in the case of a closed pipe. This is also evident if we remember that the reflection at the open end results from the suddenly increased freedom given the vibrating air, and is therefore an antinode. Three cases are shown in Fig. 53. In (a) the pipe is "speaking" in its fundamental tone, where $\lambda = 2l/1$. In (b) it is giving its second harmonic, where $\lambda = 2l/2$, and in (c) the third harmonic, where $\lambda = 2l/3$. These give frequencies whose ratio is 1:2:3:4 and so forth. In general, $n = kV/2l$, where k may be any integer, odd or even. The effective value of l used in computing the frequencies of an open pipe is a little greater than the actual length, for the perturbations representing the antinode at the upper end reach out into the air beyond the pipe. This causes a lower pitch than would be expected, but the exact amount of this correction is not easily calculated and is different for different types and dimensions of organ pipes. In consequence of this extension of l , an open pipe does not vibrate with twice the frequency of a closed pipe of the same length, as we should expect from a comparison of Fig. 53 (a) with Fig. 52. Instead it gives a note a little below the octave.

If a pipe is blown with a gas other than air, its pitch is altered. This is because of the change in the velocity of the sound within the pipe. Thus if blown with a heavy gas like carbon dioxide, its pitch falls, because with a lower value of V , and a wave length fixed by l , the frequency falls proportionately with the velocity. Similarly a gas lighter than air, like illuminating gas, carries the sound with a higher velocity, and the pitch rises in proportion.

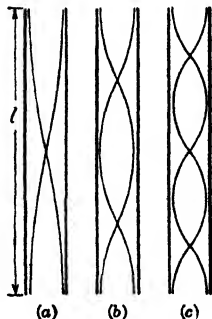


Fig. 53.

It is possible to produce standing waves in a liquid column enclosed in a tube, by setting it in vibration at one end. This arrangement also becomes a source of sound, but has no practical applications.

383. Reed pipes. The "reeds" of an organ are not true organ pipes such as were described in the last section. As has been explained, a reed set vibrating by an air blast may be the source of standing waves in a partly enclosed chamber or tube, but the reed, unlike a vibrating jet of air, has a certain frequency of its own. This necessitates tuning between the reed and its pipe, so that the latter may have the same natural period as the reed's fundamental, or one of its overtones.

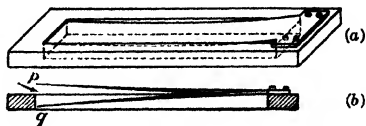


Fig. 54.

There are two kinds of organ reeds, one of which is "free," and the other a "striking" reed. They are both thin metallic tongues held rigid at one end, and vibrating at the other. A free reed swings back and forth like a door through an opening which it nearly closes when at rest, but allows the air blast to pass at either end of its swing. In Fig. 54 (a) is shown this arrangement in perspective with dotted lines to indicate the opening in the block which supports the reed. Fig. 54 (b) is a section showing the reed in its two extreme positions, *p* and *q*, which permit the air blast, indicated by the arrow, to pass through. The free reed came into use only toward the end of the 18th century, and is now used in the melodeon, accordion, and harmonica.

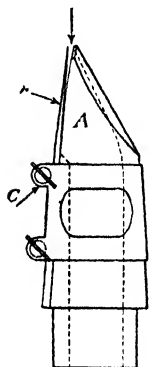


Fig. 55.

The striking reed dates from high antiquity. In this form the reed is larger than the opening, and closes it once in each period of vibration. This gives the resulting sound a very different tone color, of a more "reedy" quality than that given by the free reed.

The striking reed is used in some stops of the organ, but it is best known in such instruments as the clarinet and saxophone. The mouthpiece of such instruments is shown in Fig. 55 where the reed *r* is held against the mouthpiece by the ligature *C*, and the air is forced through the narrow opening between the reed and a chamber *A*, which the reed almost closes.

384. Vibrating membranes. The human voice is produced by forcing air between two stretched membranes, inclined in a V to each other, and known as the vocal cords. They are about three quarters

of an inch long in men, and half an inch in women. Their free edges at the vertex of the V come close together, leaving a narrow slit, and when air is forced against them, vibrations are set up which tend to close the slit periodically, thus interrupting the blast as its pressure drives the edges together. This may be imitated by a device shown in Fig. 56.

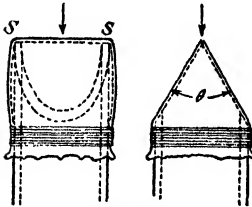


Fig. 56.

A wooden tube is beveled at one end so that the edges are inclined at some angle θ . Then if a thin sheet of rubber is fastened over it and a slit cut across the edge ss , an air blast in the direction of the arrow will cause it to vibrate. The pitch thus produced depends entirely upon the length of the tube, for the stretched rubber

membrane has no natural frequency. The vibrations of the vocal cords are controlled in various ways, such as by increased thickening near the vibrating edges (thus lowering their pitch), or by allowing only relatively small portions to vibrate, as in singing falsetto. In this way their frequency of vibration is altered without any change in the length of the air column with which they communicate, and they attain a range of a little over two octaves for ordinary voices.

The tone color of the human voice is more varied than that of any other musical instrument. This is due to the various cavities of the mouth and nose whose form and size are subject to more or less conscious control. Thus, by resonance, we impose overtones of different intensity and pitch upon the fundamental vibrations of the vocal cords.

Double-reed instruments like the oboe (*hautbois*) and bassoon really operate like the tube and membrane described above, though the vibrating edges are made of cane instead of a flexible membrane. In Fig. 57 is shown such an arrangement of two nearly parallel reeds whose edges are brought close together, forming the narrow slit ab through which air is forced by the performer. This air blast causes them to vibrate while alternately closing and opening the slit.

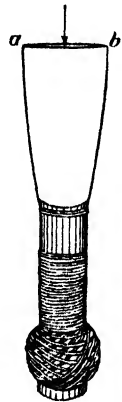


Fig. 57.

In the brass instruments, cornet, French horn, trombone, and others, the tone is produced in a similar manner, but in this case the vibrating "membranes" are the lips of the performer, who forces them tightly stretched against a cup-shaped mouthpiece. He then blows between them and sets up the vibration whose frequency is determined by the length of the air column beyond, and the number

of nodes formed there, which are controlled by the performer. In these instruments, after the little-used fundamental, the next five harmonics are produced by increasing the tension of the lips and blowing harder. Horn players produce even more overtones. Intermediate notes are made by varying the length of the air column.

385. Kundt's tube. Although not a musical instrument, this useful piece of apparatus has certain features in common with the closed



Fig. 58.

organ pipe. Unlike any musical instrument, the vibrations of the air column are set up by the longitudinal vibrations of a metal rod. The arrangement is shown in Fig. 58, where the rod AC is clamped at its center with a cap M tightly closing the end of a glass tube. At one end of the rod is a light piston, A , of hard rubber loosely fitting the tube so that its vibrations may be unimpeded. Another piston, B , fits the tube snugly, but can be moved in or out, so as to vary the length of the air column between itself and A . Finally, cork dust is strewn as evenly as possible between the two pistons. If then C is stroked with a leather washer covered with rosin, and caused to vibrate longitudinally with a node at M , it emits a high-pitched tone, and the air between A and B is made to vibrate in unison with it.

There is always a node at B , but the condition at A when resonance is established is somewhat uncertain, though for maximum resonance it must be rather near a nodal point also. The plunger is then adjusted while the rod is being stroked, until the cork filings indicate



Fig. 59.

the formation of nodes and loops. At the antinodal regions the dust is thrown violently into the air when the tone is produced, and then settles down into the characteristic Kundt's figures. These depend upon the accuracy of tuning and other circumstances. Usually the pattern is as in Fig. 59, where the nodal points are indicated by small circles and the loops by a succession of transverse ridges longest at the antinodes. However, with very perfect resonance the dust may

collect in heaps at the nodes, leaving the antinodes bare, because of the vigor with which it is thrown about by the air vibrations in the antinodal regions.

The wave length of the sound in the air of the tube is equal to the distance from one node to the second from it, or the corresponding distance between antinodes. If the air column is long enough to have several standing waves, a series of such distances can be observed and λ measured with some precision. Then if λ is known, and the velocity of sound in air calculated for the actual temperature, we may determine the velocity of sound in the metal of the rod as follows: Let V_a be the velocity of sound in air, and λ_a its wave length. Let V_r be the velocity in the rod and λ_r its wave length. This latter quantity is twice the length of the rod, because the vibrations of the rod are the longitudinal vibrations of sound, and its ends are antinodes with a single node between them, and so vibrate in opposite phase. Then $V_a = n\lambda_a$ and $V_r = n\lambda_r$. But n , the frequency, is obviously the same in both bodies; therefore dividing one equation by the other, we obtain

$$V_r = V_a \frac{\lambda_r}{\lambda_a}, \quad (1)$$

where V_a , λ_a , and λ_r are known.

It is also possible to determine the velocity of sound in a gas different from air, by introducing it through one of the stopcocks *a* or *b* (Fig. 58), while the other is opened to allow the air to escape. Then by repeating the experiment, the new wave length λ_g produced at resonance is measured. Then $V_g = n\lambda_g$, and $V_a = n\lambda_a$; whence

$$V_g = V_a \frac{\lambda_g}{\lambda_a}. \quad (2)$$

This result may now be used to determine the ratio of the specific heats of the gas in question, using Laplace's equation

$$V_g = \sqrt{\gamma \frac{p}{d}}. \quad (3)$$

If the pressure p is that of one atmosphere, we have only to read the barometer and convert the result into dynes per square centimeter, while the density d may be found by the effusion method described in Article 158. With V_g , p , and d known, the ratio γ of the specific heats may be calculated.

SUPPLEMENTARY READING

- C. F. Eyring, *A Survey Course in Physics* (Chap. 11), Prentice-Hall, 1936.
F. R. Watson, *Sound* (Chap. 10), Wiley, 1935.
J. W. Capstick, *Sound* (Chap. 10), Cambridge University Press, 1927.
E. G. Richardson, *Sound* (Chap. 7), Arnold, London, 1927.
A. B. Wood, *A Textbook of Sound* (Section 2), Macmillan, 1932.

PROBLEMS†

1. In the apparatus shown in Fig. 50, on lowering the container, resonance is first established when the air column is 52 cm long. The temperature is 22° C. What is the frequency of the tuning fork? *Ans.* 166 v.p.s.
2. How long should the air column be in Fig. 50 if a fork whose frequency is 384 v.p.s. is to establish one node between itself and the water, at 22° C? *Ans.* 67.5 cm.
3. An open organ pipe whose effective length is 90 cm is blown with air at a temperature of 20° C. What is the pitch of the second overtone? *Ans.* 573.6 v.p.s.
4. If the pipe in Problem 3 is blown with hydrogen gas whose density at 0° C and atmospheric pressure is 9×10^{-5} , what are the velocity of sound in the pipe at 20°, and the frequency of the second overtone? ($\gamma = 1.42$, $\alpha = 0.00366$). *Ans.* 1309.9 m/sec.; 2183 v.p.s.
5. In Kundt's tube, filled with air at 20° C, the dust piles average 6 cm apart, and the steel rod is 90 cm long. What is the observed velocity of sound in steel? *Ans.* 5160 m/sec.
6. When illuminating gas is substituted for air in the tube of Problem 5, the piles are 16 cm apart. What is the velocity of sound in the gas? What is the ratio of its specific heats, if its density is found to be 6.4×10^{-4} ? *Ans.* 459 m/sec.; 1.33.

† In this group of problems use 332 m/sec. for the velocity of sound at 0° C.

PART IV

LIGHT

CHAPTER 30

Production, Propagation, and Perception

386. Nature of light. Radiant energy that affects the retina of the eye is known as light. It consists of transverse electromagnetic vibrations whose wave lengths lie between 0.76 microns (10^{-4} cm), and 0.39 microns, approximately. In the phraseology of music, this is not quite an octave, and is only a very small portion of the great range of wave lengths of the various known types of radiation.

A luminous source may emit radiation having wave lengths both shorter and longer than the limits named above. These are known as the ultraviolet and infrared portions of the spectrum and, although quite invisible, are frequently referred to as "light," a term that in this case is really a misnomer.

387. Sources of light. When a solid is heated, its temperature rises, and when it is hot enough, it begins to emit visible radiation. An ideal black body begins to glow distinctly at about 500°C , and is "white hot" at around 1200°C . In general, other bodies require somewhat higher temperatures in order to reach the same degree of brightness. Most liquids are vaporized before they reach a temperature high enough to emit light, but molten metals are an exception. Gases may be made to emit light by a discharge of electricity passing through them. Ordinary flames, strictly speaking, are not luminous gases. The fact that they emit a continuous spectrum indicates, as we shall see, that most of their light is due to incandescent particles of solid matter.

Neither temperature nor electrical discharge is essential to the production of light. Chemical transformations may produce it with almost no rise of temperature. The cold light of phosphorescent substances such as decaying wood and certain fungi are illustrations. The firefly and glowworm also produce a cold light, but in this case electricity probably plays a part as well as chemical changes.

388. The measure of luminous intensity. The intensity of a source of light is most commonly expressed in terms of **candle power**. A standard candle, now rarely used, was made of spermaceti (wax obtained from sperm-whale oil), and, when equipped with a specified

wick, was supposed to burn 120 grains an hour in order to develop its rated intensity. Other standards are the pentane lamp burning the hydrocarbon of that name at a specified rate, and yielding a little more than 10 candle power; also the Hefner-Alteneck unit, which burns amyl acetate in a prescribed manner and develops 0.9 candle power.

Today, however, the incandescent lamp has practically replaced these units, whose performance was at best uncertain, and much affected by atmospheric conditions. A standard incandescent lamp gives a very constant amount of light in a particular direction at a specified voltage and current. These "secondary standards" are rated in candle power, so that a comparison may always be made with an unknown source in terms of that fundamental unit. This is still the basis of comparison in most countries, though in Germany the *hefner* is much used instead.

If the comparison of an unknown source with a 16-candle-power secondary standard shows that it is three times as intense, it is then said to have 48 candle power, although it was not compared with a candle at all.

389. Conical intensity. Let us consider the amount of luminous "flux," F , that streams through a cone at whose vertex is a luminous source of one candle power. If the cone encloses a unit spherical angle, or **ster-radian**, the amount of this energy is known as a **lumen**. The unit spherical angle is analogous to a radian because the area of a sphere of unit radius is $4\pi\text{cm}^2$, and there are 4π ster-radians about a point in space, just as there are 2π radians about a point in a plane. As the point source, supposed to be at the vertex of the cone, has a uniform intensity of one candle, it develops a total luminous flux of 4π lumens. A lumen may be similarly defined with respect to a hefner placed at the vertex of the unit cone. It has then a value of 0.9 of the lumen based upon a candle power.

390. Illumination. Illumination is a measure of sectional intensity, just as luminous flux is a measure of conical intensity. It is defined as the amount of luminous flux which falls on the unit area of a surface, normal to its path. It is measured in foot candles or meter candles, the latter unit being termed a **lux**, and is the illumination of a surface one meter distant from a standard candle, while a **foot candle** is the illumination one foot away. Average indoor daylight is of the order of 100 foot candles. Zenith sunlight is about 96,000 foot candles. One thousand foot candles is not too bright for reading, for we often have even stronger illumination when reading out of doors.

But one can read fairly comfortably with only 5 foot candles of illumination, though this involves unnecessary waste of nervous energy, as Luckiesh† has demonstrated.

If F represents the total luminous flux emitted by a luminous point source of C candle power, then the illumination I of a spherical surface surrounding it is given by $I = F/A$, where A is the area of the sphere. But C candle power emits a flux of $4\pi C$ lumens, and $A = 4\pi d^2$, where d is the radius of the sphere; therefore

$$I = 4\pi C / 4\pi d^2 = C/d^2, \quad (1)$$

where I is given in foot candles if d is measured in feet; or in luxes, if d is measured in meters. Equation (1) expresses the "inverse square law" of all forms of radiation. To be valid, the source must be a point, or a spherical surface whose center is equivalent to the point. But as W. S. Franklin‡ has pointed out, with sources other than points or spheres, the error incurred in using the inverse square law is not more than 0.2 per cent "for all cases in which the maximum dimension of the luminous source does not exceed one tenth of the distance from the illuminated surface."

If the surface is not normal to the direction of the luminous flux, but inclined so that a normal to it makes an angle θ with the beam of light, as in Fig. 1, then the angle must be considered in calculating the illumination. This is proportional to the cosine of θ , because in the case of oblique incidence, the same luminous flux is spread over an area $1/\cos \theta$ times greater than if the light were incident normally. Therefore (1) becomes

$$I = C \cos \theta / d^2. \quad (2)$$

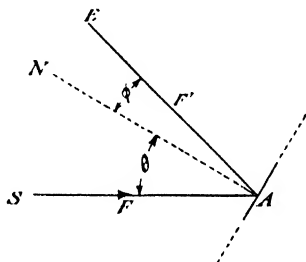


Fig. 1.

391. Brightness. When a white object is illuminated, it is said to be bright. The flame of a candle is also bright, but on its own account. Evidently brightness is a quality we perceive in a body, and it may be due either to the illumination it receives, or to its own luminosity. If the surface is not self-luminous, its brightness varies directly as its illumination, other things being equal. Two candles, shining on a piece of paper a foot away, make it twice as bright as would one candle similarly situated. Two candles give twice as much

† Matthew Luckiesh, *Light and Work*, Van Nostrand, 1924.

‡ Franklin and Grantham, *General Physics*, Franklin & Charles, 1930.

illumination as one, and twice as much illumination results in twice as much brightness on a given surface. Therefore, comparing brightness under similar conditions enables us to compare illuminations, and then the sources which caused the illuminations.

In general, the brightness of an illuminated surface depends not only upon its texture and color, but upon the angle at which we look at it. In Fig. 1, if the source is viewed along the normal N , the brightness is simply the reflected flux F' divided by the area A that reflected it. If the brightness of this same area is viewed from E , the flux reflected by A in this direction is less than before. But now we must divide, not by A , but by its *projected area*, A' , which is $A \cos \phi$. Thus we obtain the general expression for brightness

$$B = F'/A \cos \phi. \quad (1)$$

If a surface appears equally bright from all directions, as is nearly the case with plaster of Paris, it is said to be *perfectly diffusing*. Then B is constant and the flux F' from a given area A varies as the cosine of the angle ϕ . Or, since $F = IA$, $F' = I'A \cos \phi$, where I' is the intensity of the reflected light. Then (1) becomes $B = I'$, meaning that in the case of a perfectly diffusing surface, the brightness is numerically equal to the reflected intensity. The relation $B = F'/A \cos \phi$, and that relating to illumination ((2), Article 390), were discovered by J.H. Lambert (1728–1777), an Alsatian philosopher who was a pioneer in the art of photometry. The **lambert**, which is the unit of brightness, is therefore appropriately named. It is defined as the brightness of a surface which emits one lumen per square centimeter of *projected area*. This is extremely bright, so that a smaller unit, the millilambert (0.001 lambert), is much used in practice. On an ideal white surface, five foot candles would give a brightness of 5.4 millilamberts, which may be regarded as a minimum brightness for close work. If the material on which the light falls is dark, more illumination is needed to yield this required minimum of brightness. Strictly speaking, then, *brightness* should be specified in a lighting contract, because that is the final result we are interested in, whereas illumination is only the means by which brightness is achieved.

The brightness of a luminous *source* depends largely on its temperature, and is quite different from the total candle power. A source of little brightness might emit a great many candle power if it had a large luminous area, while a minute source might be extremely bright and still be low in candle power. The most concentrated luminous source on the earth is the crater of an arc light, whose brightness is

about 40,000 lamberts. The filament of an ordinary incandescent lamp has a brightness of about 500 lamberts, while in the flame of an oil lamp the value falls to 5 lamberts, though its much larger area gives it a candle power comparable to that of a 30-watt electric light.

The four kinds of units defined in the preceding paragraphs are confusing, and some may seem unnecessary. But it is difficult to see how we could do without at least three units which measure:

- (1) the luminous intensity of a source of light (for example, candle power),
- (2) the intensity of emission per unit area (for example, lambert),
- (3) the intensity of reception per unit area (for example, foot candle).

392. Photometry. The eye is quite incapable of comparing two luminous sources with any accuracy, but it can compare the brightness of adjacent and similar surfaces with remarkable precision. In fact, a difference of one part in 150 may be detected, provided the colors are nearly the same. As explained above, this enables us to compare the illuminations of the surfaces and so the sources of those illuminations.

A very simple device for making such a comparison was devised by Count Rumford. A screen of dull-surfaced paper is placed so as to receive the light from the two sources to be compared, in such a way as to be equally inclined to both beams. A rod R , shown in plan in Fig. 2, casts two shadows, a and b , on the screen AB , as indicated in its elevation $A'B'$. If the distance between rod and screen is properly adjusted, the shadows may be made to touch each other without overlapping. The area a is shielded from S_2 , but receives light from S_1 , while b is shielded from S_1 and is illuminated by S_2 . If the two shadows are made equally dark by varying the distances of S_1 and S_2 from the screen, then the illuminations are equal, and using equation (2), Article 390, we have

$$I_a = C_1 \cos \phi_1 / d_1^2 = I_b = C_2 \cos \phi_2 / d_2^2.$$

If ϕ_1 is made equal to ϕ_2 , then the candle powers of the two sources are to each other as the squares of the distances, or $C_1/C_2 = d_1^2/d_2^2$, from which either may be determined if the other is known.

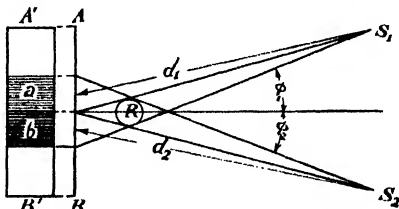


Fig. 2.

393. The Bunsen and Lummer-Brodhun photometers. In both of these photometers, there is a constant distance between the two lamps to be compared. A movable screen between them is adjusted until both surfaces are equally bright, as indicated in Fig. 3. In

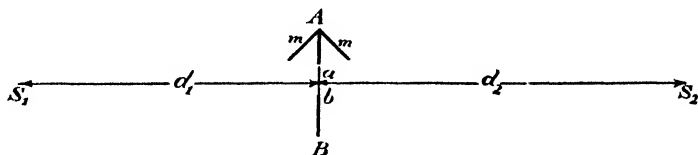


Fig. 3.

Bunsen's form, the screen is made of thick paper with a spot of grease at the center, or an inset of translucent tissue paper. When the brightness is the same on both sides of the screen, as seen in the mirrors *mm*, the thinner or greased area *ab* is equally bright with its surroundings which transmit no light; otherwise it is not. This is because the translucent spot transmits, say, one quarter of the light it receives from *S*₁ on its left face, and reflects three quarters. Its brightness on its left face from this source is therefore only three quarters as great as that of its opaque surroundings. To this amount must be added the one quarter due to *S*₂ received on its other face and transmitted. Then on the left its total effective illumination is

given by
$$I_L = \frac{3C_1}{4d_1^2} + \frac{C_2}{4d_2^2},$$

and on the right by
$$I_R = \frac{3C_2}{4d_2^2} + \frac{C_1}{4d_1^2}.$$

If the screen is adjusted so that $C_1/d_1^2 = C_2/d_2^2$, then $I_L = I_R$ and the grease spot is equally bright with its surroundings, which in turn are equally bright themselves when viewed from either side. When this balance is effected, the candle power of either source is readily computed, if the other is known, from $C_1/C_2 = d_1^2/d_2^2$.

In the Lummer-Brodhun photometer, shown in Fig. 4, the screen *A* is of magnesium oxide or plaster of Paris, is wholly opaque, and gives a nearly perfectly diffuse reflection. The brightnesses of its two surfaces are compared by a system of prisms acting as mirrors, so that the surface illuminated by *S*₂ is seen as a ring of light reflected first by the mirror *m*₂, and then by the reflecting ring *m*₃. This ring of light surrounds the light from the surface of *A* illuminated by *S*₁ and reflected by mirror *m*₁. The reflected light passes through a hole in

m_3 and forms the central disc in the field of vision, shown lighter than the outer ring in the diagram. The photometer is balanced to make these areas equally bright by sliding the system of screen, mirrors (really prisms), and viewing telescope (not shown), along a track on which a scale is laid off in centimeters, so that d_1 and d_2 are easily measured. Then the illumination of the two faces of the screen must be the same and the usual photometer formula $C_1/C_2 = d_1^2/d_2^2$ may be applied.

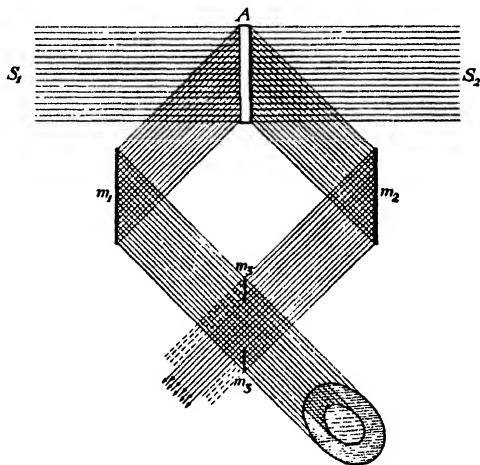


Fig. 4.

394. Rays of light. As was explained in Article 322, a ray is any line drawn perpendicular to the front of a spherical wave. If the medium is isotropic, a point or a spherical source emits a spherical wave front, and the rays are radii of that sphere. Although a pure fiction, rays are extremely useful in studying the more obvious phenomena of light, but the ray construction is quite inadequate when we are concerned with such phenomena as interference and diffraction, as will be seen further on.

When light is treated as a ray phenomenon, the problems considered are purely geometrical, and whenever this procedure is sufficiently exact to be permissible, the subject matter is known as *geometrical optics*. However, in problems for which the simple geometrical method is inadequate, we must use wave construction. Such aspects of light come under the head of *physical optics*, because they are explained only by an understanding of the actual physical properties of matter, and by regarding light as propagated by waves and not by rays.

395. The propagation of light. It was not known to the ancients that light has a finite speed. This is hardly to be wondered at, since it travels so fast that our senses are quite inadequate to perceive any lapse of time, for example, between a flash of light near by and its reflection from a distant mirror. Moreover, there seems to have been

a curious idea in early times that seeing things consisted in some form of intangible contact reaching out from the eyes to the object seen. An old woodcut illustrates this naïve idea by having arrow-like lines projecting from the eyes of an individual looking fixedly at something, instead of having the arrows entering his eyes as they should. This conception still lingers in our daily language when we speak of trying to "pierce the darkness" with our eyes.

The philosophers of the middle ages developed a corpuscular theory of light. They imagined that the luminous object emitted a stream of particles which impinged upon the eye and gave rise to the sensation of light. These corpuscles were supposed to travel at a very high speed which, however, was not considered infinite. Another and later theory was due to the eminent French philosopher Descartes (1596–1650), who explained the propagation of light as a kind of rotary motion in a medium of infinite elasticity which transmitted it with infinite velocity as a steel shaft transmits energy in a factory. But Newton, born eight years before Descartes' death, adhered to the corpuscular theory, although Huygens had already advocated a wave hypothesis. So great was Newton's prestige that it was not until the early part of the last century that the conclusive experiments of Fresnel† finally put an end to the "corporeity of light," as Newton called his theory. This was of course a great step forward, but it must be admitted that the corpuscular theory under Newton's able manipulation was made to account for all the more usual optical phenomena, though it failed to explain those already enumerated as belonging to the realm of physical optics.

• **396. Römer's determination of the velocity of light.** In 1675, Ole Römer, a Danish astronomer, read a paper before the French Academy of Sciences in Paris, in which he announced a calculation of the velocity of light based upon the known irregularities in the interval between successive eclipses of the first (innermost) moon of Jupiter by that planet. The method is as follows: When Jupiter is in opposition, as indicated by *J* and *E* in Fig. 5, the average interval between eclipses is 42 hours, 28 minutes, and 36 seconds. As the earth moves farther away because of its greater orbital velocity, this interval steadily increases. It reaches a maximum of about 15 seconds longer than the average at some point *a* when its motion is directly away from Jupiter, which is then at *b*. After this the interval

† Augustin Jean Fresnel (1788–1827), a French engineer and man of science. His famous "Mémorial on the Diffraction of Light" was presented to the French Academy in 1818.

between eclipses gradually grows less until, when the earth is at E' , nearly opposite to E , it once more has the average value.

During the second half year, the interval decreases, reaching a minimum at c , when the speed of approach to Jupiter is greatest. This is followed by a gradual recovery of the average value when

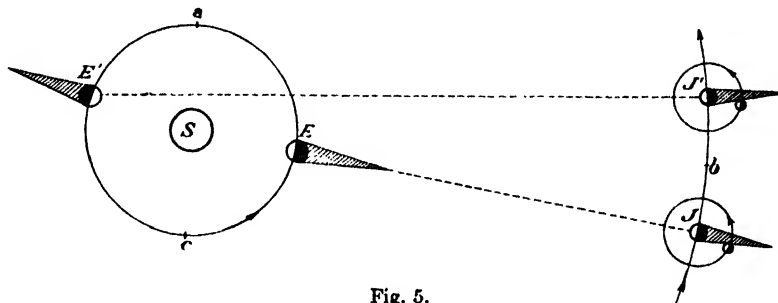


Fig. 5.

Jupiter is once more in opposition, with the earth advanced about a month on the second year, because Jupiter's year is more than twelve of ours.

The meaning of these changes in the observed interval is that we are dealing with a sort of Doppler effect. Jupiter's moon sends us signals at a nearly constant rate. The time interval between the reception of two successive signals is the same regardless of our distance from their source, if that distance is constant. Though they take time to reach us, each of two signals marking an interval is equally delayed, just as the sounds from a "minute gun" would reach us a minute apart whether we were ten feet away or ten miles. But when the distance changes during the time between signals, one signal is delayed more than the other, and the interval is altered. There are 112 such intervals between the two positions E and E' . If we measure the interval at E , when the earth is neither approaching nor receding from Jupiter, and multiply by 112, we can calculate when the eclipses should be seen at E' if light had infinite speed. But because of the increasing length of the intervals as the earth receded from Jupiter (or shortening during approach) Römer found that at E' the eclipses would be 996.4 seconds late, though the interval between successive eclipses must be the same as at E .

Taking the mean diameter of the earth's orbit as 186 million miles, Römer reasoned that it took light 996.4 seconds to travel this distance, and that its velocity was therefore $186 \times 10^6 / 996.4 = 187,000$ miles per second. Later and more accurate measurements resulted

in 1001.6 seconds for the total time, with a probable error of one second. This means that the velocity of light is nearly 186,000 miles per second, or 3×10^{10} centimeters per second.

397. Bradley's method. Römer's determination of the velocity of light was not generally accepted until, in 1729, James Bradley, an English astronomer, obtained a value of 186,400 miles per second by observing the "aberration" of the stars. This is an apparent displacement of their positions due to the earth's motion, similar to the apparent change in the direction of falling rain when we are moving rapidly through it. The change depends upon both the observer's speed and that of the rain, and if the observer's speed is known, the speed of the rain drops may be calculated from the apparent change in their direction.

398. Fizeau's method. In 1849, Hippolyte Fizeau, a French natural philosopher, made what may be called the first laboratory measurement of the velocity of light. This method consisted in sending a beam of light across the edge of a rapidly revolving toothed wheel so that it was flashed intermittently through the spaces between the teeth. This was set up in the village of Suresnes, near St. Cloud, outside of Paris, and the flashes were reflected back again by a mirror on Montmartre in the city, a distance of 8.633 kilometers each way. When the wheel was at rest, the light, passing through a slot, was returned along the same path by a system of lenses which

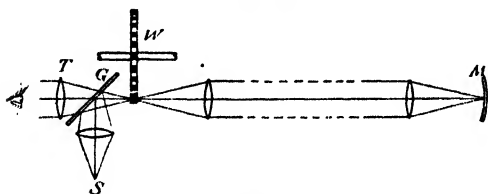


Fig. 6.

concentrated the light on the same opening through which it had originally passed, as shown in Fig. 6. The outgoing light from *S* was partly reflected by the surface of a sheet of

plate glass *G* inclined to 45° , and on its return it was partly transmitted by the same plate to the telescope *T*, where it appeared like a star. When the wheel was rotated with increasing speed the star became gradually fainter until it was wholly eclipsed. This occurred when the time required for a slot to replace an adjacent tooth was just equal to the time required for the light to travel out and back, a total distance of 17.266 kilometers. At double this speed the star again reached a maximum of brightness. At triple speed it was again extinguished, and so on. The first eclipse occurred with a speed of 12.6 r.p.s. Then, knowing the angular separation of a tooth and a slot, it was easy to

calculate the time required for the light to travel out and back. This was $1/18,144$ of a second, so the observed velocity was $18,144 \times 17.266 = 3.13 \times 10^5$ km./sec., or 1.95×10^5 mi./sec., a result which is about 5 per cent too large.

399. Foucault's method. A much more reliable method for measuring the velocity of light is due to the French physicist, Leon Foucault (1819–1868), though his method has been much improved upon by Newcomb and Michelson, both Americans.

Foucault's original experiment was performed in 1850, and consisted in causing a beam of light to be reflected from a rapidly revolving plane mirror at the center of curvature of a concave mirror a few meters away. The returning light met the revolving mirror at a different angle from its first reflection, and so instead of being returned along its original path, it suffered a displacement which varied with the mirror's angular velocity.

In 1878, Michelson repeated the experiment, using a plane mirror at a distant station, and so was able greatly to increase the length of the path over which the light traveled. The essentials of the apparatus are shown in Fig. 7. Light from the source S (a narrow illuminated slit) is reflected from the rotating mirror m , and its divergent rays are made into a parallel pencil by the lens L . This is reflected by the mirror M some 600 meters distant and returned upon its own path, being concentrated by L so that after reflection it is brought to a focus at S . Now if m is revolving about the axis a in the sense indicated by the arrow, it will send out one flash per revolution each time it passes through the position indicated by the heavy line, and will have turned through the angle θ before the flash has returned from M . Thus the reflected light from each flash is brought to a focus at E instead of at S , and the angle SaE , measured by the ratio of the arc SE to the radius r , is equal to 2θ , as is proved in Article 408. But if θ and the angular velocity of the mirror are known, the elapsed time is easily computed. In this way Michelson obtained a value of 2.999×10^{10} cm./sec., or 186,300 miles per second for the velocity of light.

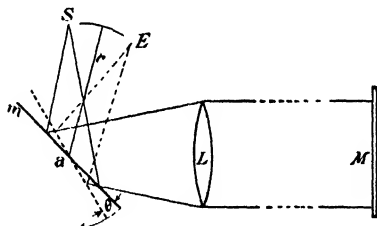


Fig. 7.

400. Michelson's method. Acting on a suggestion made by Simon Newcomb, Michelson, during the years 1924–27, made use of a re-

volving octagonal mirror and a so-called null method, in which the reflected beam was not deviated. In Fig. 8, the general principles involved are shown diagrammatically and greatly simplified. Light from a source S was reflected from a face of the octagonal mirror M .

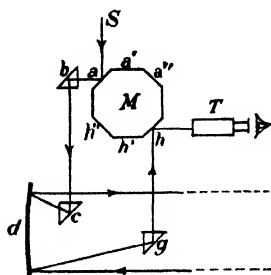


Fig. 8.

It was then reflected by the prism b (b , c , and g are totally reflecting prisms) to c , actually at the principal focus of the concave mirror d , which sent a parallel beam to a similar mirror, e , 22 miles away, from Mt.

Wilson to Mt. San

Antonio in Southern California. This beam formed an image on a small concave mirror f at the principal focus of e , and was thus returned over the same path to d . Then the beam was focused on another prism g , so inclined as to return it to the mirror face h parallel to a . From h the beam was reflected into the telescope T . If the octagonal mirror were at rest in the position shown, the light would be seen in the telescope. But when set in rotation, the light disappeared until the octagon was rotating at the exact speed needed to bring h' into the position just occupied by h during the time taken by the light to travel the total path $abcdefedgh$. Thus each flash sent out from a , a' , and so forth, was picked up by the faces h' , h'' , and so forth, next to those which would reflect it if the mirror were at rest. The mirror must rotate 45° during this interval, and if the critical speed needed to restore the light, and the various distances are known, the velocity of light may be calculated.

During the years 1930-33, Michelson (until his death in 1931), Pease, and Pearson made similar measurements of even higher precision in a three-foot tube a mile long and exhausted to a pressure of half a millimeter of mercury. The rotating mirror had 32 faces, and the light traveled back and forth nine times between the mirrors at the ends of the tube during the time of replacement of one of the 32 faces by the next one. The resulting value of the velocity of light in a vacuum, obtained in February, 1933, is 299,774 kilometers, or 186,271 miles, per second.

401. The visible universe. Our knowledge of the visible universe depends upon our perception of relative brightness, of the apparent

size and shape of objects, of their relative positions, and of their color. A combination of these four aspects, combined with knowledge derived through the sense of touch in handling visible objects, gives us a mental picture of things at a distance which corresponds fairly accurately to their actual size and shape and relative position. Relative brightness is an important guide in judging distance, as is also color, for we have learned to associate the blue tint, called "atmosphere" by painters, with distance. Outline and shading are clues to shape, since we can differentiate between a cube and a sphere without touching them, while the pattern made by a group of objects tells us much about their relative positions.

402. Binocular vision. Perhaps our most valuable means of judging relative distance is derived from the fact that our two eyes give us two different pictures. If a vertical rod a few feet away is viewed with one eye only, it has a definite position against the more distant background, but when seen with the other eye this position changes. Thus a tree trunk seen in plan at T (Fig. 9) appears at a against a distant mountain when viewed by the right eye, E_2 , alone, and at b by the left eye, E_1 , alone. The two pictures, E_1 and E_2 , are transmitted to the brain, which interprets the combination from experience ac-

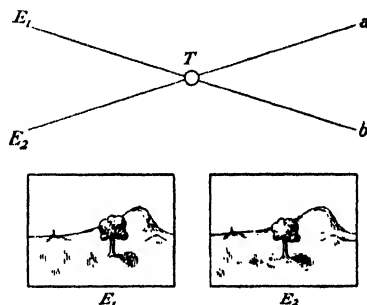


Fig. 9.

quired in our earliest infancy, as meaning that the tree is nearer than the mountain. A long *horizontal* rod or wire however gives no such clue, because our eyes are placed horizontally. This is the reason that it is difficult to judge how near we are to a wire stretched across our field of vision. To judge its distance, the observer should hold his head horizontally so that one eye will be above the other.

403. The stereoscope. This is a device for creating the illusion of "depth" (distance in the line of sight) from two photographs taken side by side, like the two ocular images of ordinary vision. These photographs are viewed through prisms or prismatic lenses which enable each eye to observe the proper photograph independently of the other, so the result is similar to binocular vision of the original scene.

A single photograph is a one-eyed picture, and when seen by two eyes, we *know* it is flat because both visual pictures are identical, and the only illusion of depth is due to perspective and relative brightness.

or *chiaroscuro*. Hence a one-eyed picture should really be viewed by one eye only, because then we do not miss the absence of the stereovision (solid vision), and are more easily tricked by other considerations into seeing depth.

In the stereoscope, the observer sees in the two pictures (each viewed by one eye) just what he would have seen with both eyes open at the place where the photograph was taken, provided his eye is where the lens was with respect to the photograph, and provided the distance between the lenses of the stereoscope camera is the same as the distance between the eyes. Unfortunately, this latter condition seldom occurs. The lenses are usually farther apart than the eyes, and all nearer objects are dwarfed for a reason to be explained in the next article.

404. Apparent size. "How large does the moon look?" This is a question often asked, but it is quite meaningless. It all depends upon how far off the object is with which we compare it. A dinner plate one hundred feet away would just about hide it, while a good-sized pea held at arm's length would do the same. The fact is that the only real meaning in this estimate of size is the angle the object subtends at the eye. Two men of the same height but at different distances subtend different angles, and if we had no means of estimating their distance from us, we should assume the more distant one to be shorter than the other. Conversely, if they are at the same distance

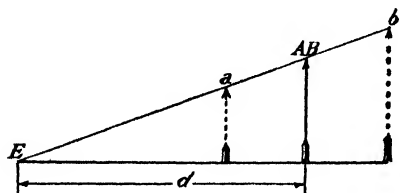


Fig. 10.

but appear to be at different distances, the one seen as nearer appears the smaller of the two. Thus if two objects, *A* and *B* in Fig. 10, are at the same distance *d* from the eye *E*, but for some reason *A* appears to be at *a* and *B* at *b*, then obviously *A* seems reduced and *B* enlarged

because both subtend the same angle at *E*. This accounts for the apparent magnification of the sun or full moon when seen near the horizon, especially if over a long stretch of open country. The intervening landscape forces us to regard the luminary as being a long way off, and our estimate of its size is enhanced. When overhead, there is nothing with which to gauge its distance, and instinctively we bring it near us, thus diminishing its apparent size.

Fog and darkness have the same effect of creating the illusion of increased distance of objects as compared to their apparent distance

when seen in a good light, and cause them to "loom up" with surprising bigness.

It is now easy to see why stereoscopic pictures reduce the apparent size of close-up objects such as people in the foreground of a landscape. The fact that the lenses of the camera are usually farther apart than the eyes, increases the displacement indicated in Fig. 9. This would also be the case if the tree were brought nearer; therefore the observer instinctively judges it to be nearer than it is, and since the angle subtended is not actually altered, he is forced to consider it as reduced in size.

SUPPLEMENTARY READING

Hardy and Perrin, *The Principles of Optics* (Chap. 13), McGraw-Hill, 1932.
J. Valasek, *Elements of Optics* (Chapters 1, 2), McGraw-Hill, 1928.
J. P. C. Southall, *Mirrors, Prisms and Lenses* (Chap. 1), Macmillan, 1933.
R. W. Wood, *Physical Optics* (Chap. 1), Macmillan, 1934.

PROBLEMS

1. How many lumens are emitted by a lamp of 16 candle power? *Ans.* 201 lumens.

2. How many lumens pass through a spherical area of 120 cm^2 having a radius of 80 cm, when a point source of 50 candle power is at the center? *Ans.* 0.94 lumen.

3. How many lumens fall on 2 square feet of a screen whose illumination is 3 foot candles? *Ans.* 6 lumens.

4. What is the illumination of a surface 40 cm from a 16 c.p. point source, if the surface is normal to the luminous flux? *Ans.* 100 lux.

5. What is the illumination of a surface 60 cm from a 40 c.p. point source if the normal to the surface makes an angle of 60° with the flux? *Ans.* 55.6 lux.

6. A plaster of Paris surface is one meter from a 40 c.p. point source, and its normal is inclined 45° to the incident flux. What is its brightness as seen from any angle, assuming an absorption of 0.1 by the plaster? *Ans.* 2.56 mililamberts.

7. In a Bunsen photometer, the greased spot vanishes when the screen is 120 cm from a standard 16 c.p. lamp, and 80 cm from an oil lamp. What is the candle power of the latter? *Ans.* 7.1 c.p.

8. How far from a screen should a 27 c.p. lamp be placed to produce the same illumination as a 3 c.p. lamp at 4 ft.? *Ans.* 12 ft.

9. An electric light giving 36 c.p. and an oil lamp giving 16 c.p. are 12 ft. apart. How far from the oil lamp in the direction of the electric light is the illumination from both sources equal? *Ans.* 4.8 ft.

CHAPTER 31

Reflection

405. Laws of reflection. There are two fundamental laws of reflection. One is that the angle of incidence is equal to the angle of reflection, in accordance with Huygens' principle, as was proved in Article 321. If that principle is accepted, this law of reflection must be accepted also. However, it has also been tested experimentally in many ways, and the most delicate measurements have failed to detect any deviation from an exact equality between the two angles.

The second law states that the incident and reflected rays, as well as the normal to the reflecting surface, all lie in the same plane, called the **plane of incidence**. This also follows from Huygens' principle, and is easily demonstrated by direct observation.

406. Images in a plane mirror. An image of a point, as was shown in Article 322, is on a line drawn from that point normal to the

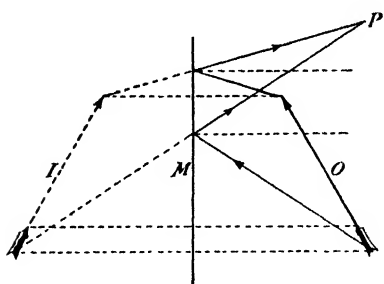


Fig. 11.

reflecting surface, and as far behind it as the object lies in front. Images of finite bodies are the totality of all the images of all their component points. Thus in Fig. 11, the object O reflected in the mirror M (seen in section) has an image I , constructed by drawing the perpendicular dotted lines, and laying off equal distances on both sides of the plane

of the mirror. This means that an observer at some point P would see the arrow exactly as if it were really at I , although the light reaching him comes by the route indicated by the solid and not by the dotted lines, these being rays reflected at the surface in accordance with the law of equal angles.

A plane mirror may be regarded as a window looking into *image space*, which is theoretically that half of the universe lying in front of the infinite plane of the reflecting surface. If a mirror is set in a nonreflecting wall, as shown in Fig. 12, it is obvious that it acts as

such a window for the point object b . But it is not quite so evident that a point a has an image a' lying above the top of the mirror. However, if it really were a window, we should look from some point c in order to see a' , and the same is true in observing the reflection of a , as is shown by the solid lines representing the course of a reflected ray. Thus, by taking a proper position, we can see reflected in the mirror any object, not hidden behind something else, which lies in object space to the right of the mirror's plane. Therefore all image space contains all points in object space, but behind the plane of the mirror.

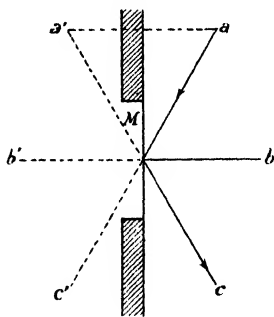


Fig. 12.

This convenient fiction of a window into image space helps to solve many simple problems of reflection. Thus it is evident that a person whose eyes are at E in Fig. 13 needs a mirror of only about half his own height in order to see his whole figure EG . His image $E'G'$ is obviously completely visible from E , and this would be equally true at any distance from the mirror, because if he steps back to the line eg , his image recedes by the same amount to $e'g'$.

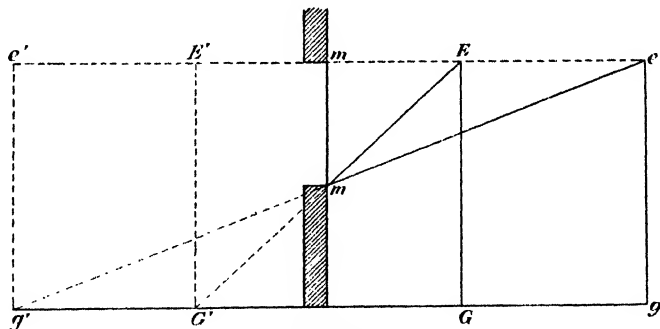


Fig. 13.

407. Perversion and inversion. Plane mirrors, when facing an object, are said to reflect a perverted image of the object, that is, turning left for right. But they do not *invert* it, turning up for down. A vertical printed page held at right angles to a vertical mirror can be seen simultaneously with its image, and in the latter the type is all reversed. However, if the mirror were held under the page and at right angles to it, an unperverted image is seen, but now it is inverted

instead. A plane mirror does not pervert and invert at the same time, as is done by concave mirrors and lenses when they form a real image (Articles 416 and 434).

When an object is parallel to a mirror, it is not so obvious why its image should be described as perverted. When you look in a mirror, your right hand is reflected on the right, and your left hand on the left, but viewed *from the mirror*, object and image are turned left for right.

408. Rotating mirrors. If a ray of light is incident on the mirror M at an angle i with the normal N as shown in Fig. 14, it is reflected

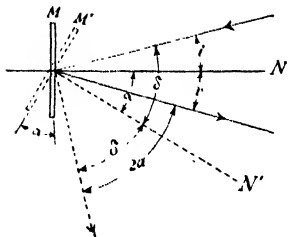


Fig. 14.

at an angle r . But $r = i$, so the angle between the two rays is $2i$. Then if the mirror is rotated through an angle α to the position M' , the normal turns through the same angle to N' , and the angle of incidence is increased by the angle α , and is now δ , or $i + \alpha$. The reflected ray makes the same angle with the new normal N' as the incident ray, so that the angle between them is 2δ , or $2(i + \alpha)$. But

this angle was originally $2i$, so the reflected ray has been turned through 2α . Therefore, when a mirror turns through any angle about an axis normal to the plane of incidence, the reflected ray turns through twice that angle.

409. Two mirrors. An object placed between two mirrors perpendicular to a common plane, and inclined at an angle θ to each other, gives rise to a series of images which lie on the circumference of a circle, whose center is in the intersection of the planes of the mirrors. In Fig. 15, the object P lies between the mirrors A and B , whose planes (perpendicular to the paper) intersect at O . The image p_1 formed in the mirror A is at the same distance from O as P is, because the triangles aOp_1 and aOP are equal. For the same reason P and the image q_1 formed in the mirror B are equidistant from O . Therefore a circle drawn through P , with O as a center, passes through p_1 and q_1 . In the same manner the image p_2 of p_1 , formed in B , lies at

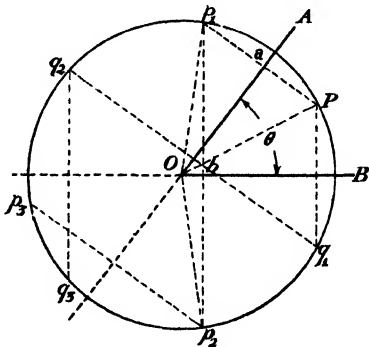


Fig. 15.

the same distance from O as p_1 , because the triangles bOp_1 and bOp_2 are equal; therefore p_2 lies in the same circle as p_1 and q_1 . In this way it is easily seen that all the images of images lie at a common distance from O , and are therefore in the same circle.

The series of images beginning with p_1 ends with p_3 formed by A , because this point lies behind the reflecting surface of B and can therefore have no image formed there. Similarly q_3 (formed by B) ends this series because it is behind the reflecting surface of A .

410. Two mirrors normal to each other. If $\theta = 90^\circ$, a case commonly met with, there are three separate images, and if $\angle POB$ is 45° , they form

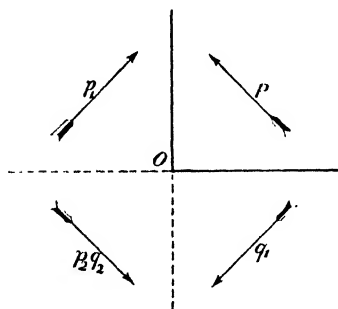


Fig. 17.

image p_2q_2 is an unfamiliar one with its left side opposite his right. He sees himself as others see him.

411. Parallel mirrors. If the angle θ between two mirrors is decreased by rotating one of the mirrors around its outer edge C (Fig. 18), the center of the image circle shifts to some point O' , its radius increases, and the curvature of the succession of images decreases until, when the mirrors lie one behind the other in a straight line, and their number is infinite.

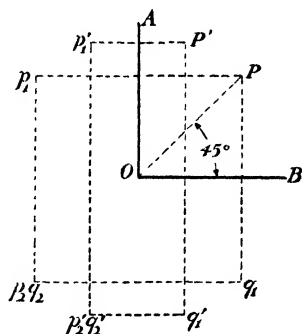


Fig. 16.

with the object the corners of a square; otherwise they form an elongated rectangle, as shown in Fig. 16. The images p_1 and q_1 of a finite object at P (Fig. 17) are perverted in the sense already explained, but the coincident images p_2 and q_2 are again perverted and therefore constitute an unperverted image of P , because when seen from O they both look alike. However, if an observer at P views himself in the mirrors, the twice-reflected

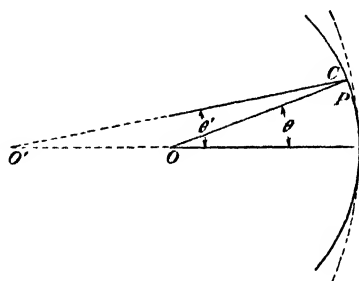


Fig. 18.

After an even number of reflections from parallel mirrors, the final ray is parallel to its original direction, as is seen in Fig. 19, where the

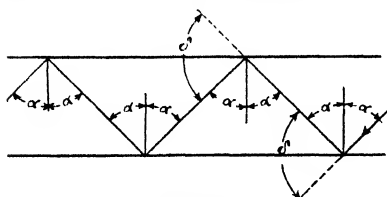


Fig. 19.

deviation δ is $180^\circ - 2\alpha$ as a result of the first reflection, but the ray is bent through the same angle in the opposite sense by the second reflection, and so on indefinitely. Thus after every two reflections its original direction is restored.

412. The sextant. One of the most valuable applications of two mirror reflections is the **sextant**. This instrument, invented in 1731 by John Hadley, a British astronomer, is used by mariners to determine the altitude of the sun; that is, the angle it makes with the horizon at the point of observation.

The essential parts of the sextant are shown in Fig. 20. A telescope is fixed rigidly to a frame which also carries a fixed scale D and a fixed mirror A , of which only the lower half is silvered. The scale is sixty degrees of the arc of a circle centered at the center of the movable mirror B . An arm R , pivoted at this point, carries B , which is thus caused to rotate about the axis of R , while the free end, equipped with a vernier, travels over the scale.

When R is set at O (the zero of the scale) the two mirrors are parallel, and if the telescope is sighted with its cross hairs intersecting the horizon as seen through the unsilvered part of A , it will be also sighted on the horizon after two reflections by the route $H'BAT$, because, as has just been shown, rays twice reflected from parallel mirrors are not altered in direction.

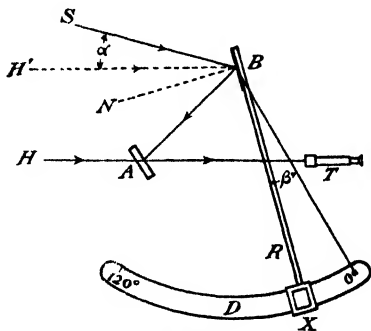


Fig. 20.

If now R is so adjusted that the sun, as seen after two reflections by the route $SBAT$, coincides with the horizon seen directly through the unsilvered half of A , then the angle α indicates its altitude. But it was proved in Article 408 that reflected rays are turned through twice the angular displacement of the mirror. Therefore β , the angle through which B was turned, is equal to half of α . The scale D is

laid off to read 120° over an arc of 60° ; therefore its divisions are really half degrees, and α is read directly without the necessity of doubling the observed angle.

413. Reflection from concave spherical mirrors. Spherical mirrors are really segments of a spherical surface, obtained by cutting off a portion of the sphere by a plane, and to be of any value as optical instruments they must be but a small portion of the total sphere. This means that they must have a small *angular aperture*, which is the angle α subtended at the center of the sphere by the segment S constituting the mirror, as shown in Fig. 21.

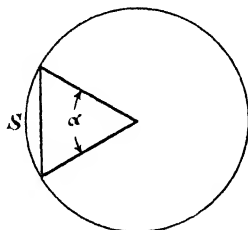


Fig. 21.

In the following discussion of mirrors we shall regard distances measured from the mirror *toward* the source of light as positive, and distances behind the mirror *away from* the source of light as negative. Then a convergent (concave) mirror has a positive radius of curvature, and a divergent (convex) mirror has a negative radius of curvature. In Fig. 22 let AVB represent the circular section of a spherical mirror of radius r whose center is at C and whose vertex is at V . Let O be a luminous point or *object* situated on the axis of the mirror, and let OV and OA represent two rays from O , the former along the axis and the latter to any point such that the arc AV subtends a *small* angle α at O . This second ray, after reflection, intersects the first at some point I . By the laws of reflection the radius CA bisects angle OAI

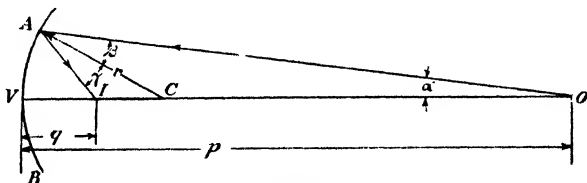


Fig. 22.

because it is normal to the surface at A , and therefore the angles β and γ are equal. The bisector of an angle of a triangle intersects the opposite side in segments proportional to the other two sides; therefore in the triangle OAI , $OC/IC = OA/IA$. But as α was supposed small, OA is very nearly equal to OV , and IA very nearly equal to IV ; therefore

$$\frac{OC}{IC} = \frac{OV}{IV} \text{ very nearly,} \quad (1)$$

for a mirror of small aperture. This equation is true for all rays from O lying within an angle small enough to justify the above assumptions; therefore I is the image of O , because all the rays diverging from O (within the prescribed limits) must intersect at I , as shown in Fig. 23.

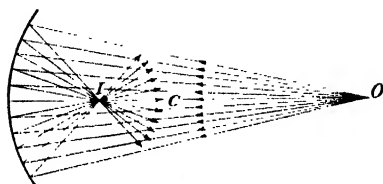


Fig. 23.

Now let p be the object distance OV , and q the image distance IV , and let $CV = r$, the radius of the sphere. Then $OC = p - r$, $IC = r - q$, and equation (1) becomes

$$\frac{p - r}{r - q} = \frac{p}{q}.$$

Clearing of fractions and combining terms, we obtain

$$qr + pr = 2pq,$$

and dividing by pqr , we have

$$\frac{1}{p} + \frac{1}{q} = \frac{2}{r}. \quad (2)$$

Equation (2) enables us to calculate any one of the three quantities when the other two are known. In general, r is given, and it is desired to find either the object or the image distance. Suppose a given concave mirror has a radius of curvature of 50 cm; then a point object placed at 75 cm from the mirror on its axis produces an image at a distance given by

$$\frac{1}{q} = \frac{2}{50} - \frac{1}{75}; \text{ hence } \frac{1}{q} = \frac{4}{150}, \text{ and } q = 37.5 \text{ cm.}$$

If p is infinite, $1/p = 0$, and $q = r/2$. Now all rays from an infinitely distant source are parallel to the axis, and have a plane wavefront; therefore plane waves are brought to a focus at a point midway between the mirror and its center of curvature. This point, to be denoted by F , is known as the **principal focus**, and its distance from the vertex of the mirror is the **focal length**, denoted by f . That is, $f = r/2$, showing that f increases with decreasing curvature of the mirror. This quantity is much used in calculations regarding both mirrors and lenses, and the principal focus is a point of great value in making graphic constructions of images.

Equation (2) may now be expressed as

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f}, \quad (3)$$

which is the most usual form of the concave mirror equation.

414. Object and image relations. If the object is at C , then $p = r$, and from (2) and (3) we obtain $q = r = 2f$. This is almost self-evident, for the rays sent out from an object at C are all radii of the sphere and therefore normals to its surface, so the reflected rays come back again to their source.

If p is less than $2f$ but greater than f , for instance, 37.5 cm as in the above problem, then q is 75 cm. Thus object and image have exchanged places. This is to be expected, because there is nothing in the geometrical method of treating optical phenomena which depends upon the *sense* of the ray's path. In reflection, for instance, the angles of incidence and reflection are interchangeable, and an object seen in a mirror may be interchanged with the observer's eye with no alteration in the diagram. In like manner we should expect that a luminous source at f would, after reflection, produce a plane wave or parallel rays with a focus at infinity. This follows from equation (3) by setting $p = f$. Then $1/q = 0$, or $q = \infty$.

Finally, if p is between f and the mirror,

$$\frac{1}{q} = \frac{1}{f} - \frac{1}{f - e},$$

where e is any positive number less than f . Evidently $1/(f - e)$ is greater than $1/f$; therefore q must be negative. This means that the image lies on the opposite side of the mirror from the object, and it is then called a **virtual image**. The rays diverge from the mirror after reflection, but if produced backward, they intersect at a point on the axis behind it, and an eye in object space sees the image in the mirror very much as it does in a plane mirror whose images are virtual also. In order to see *real images*, the eye must be farther from the mirror than the image and somewhere within the divergent beam radiating from the image, as shown in Fig. 23. From such a position the image seems really to exist in space as if independent of the mirror, but unlike a genuine object, it cannot be seen from the side or from behind.

415. Convex mirrors. The formula for convex mirrors is derived in a manner similar to that used with concave mirrors. Thus in Fig. 24, I is the intersection of the rays from O reflected at A and V , produced to determine an image behind the mirror. This is obviously

virtual because all rays diverge after reflection, as is easily seen from the diagram. The radius CA bisects the exterior angle δ of the triangle AIO ; therefore the segments of the side OI produced are

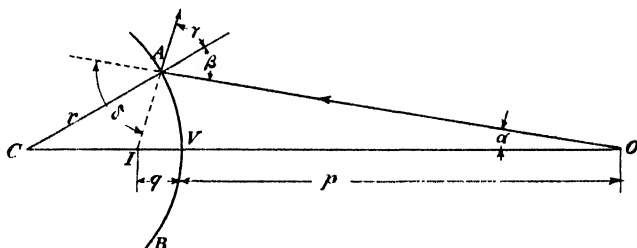


Fig. 24.

proportional to the two other sides as before, or $OC/IC = OA/IA$, and if the aperture is small, this again reduces to

$$\frac{OC}{IC} = \frac{OV}{IV} = \frac{p}{q}. \quad (1)$$

But $OC = p + r$, and $IC = r - q$.

Therefore

$$\frac{p + r}{r - q} = \frac{p}{q},$$

$$pq + qr = pr - pq,$$

and

$$pr - qr = 2pq.$$

Then dividing by pqr , and remembering that q and r are behind the mirror and negative, we obtain

$$\frac{1}{p} + \frac{1}{q} = \frac{2}{r}, \quad (2)$$

as for a concave mirror.

Thus if an object is 75 cm in front of a convex mirror of radius -50 cm, then $1/q = -2/50 - 1/75 = -4/75$. The image is 18.75 cm *behind* the mirror, as indicated by the negative sign. On the other hand, it might be stated that an object 75 cm from a mirror (convex or concave) has a virtual image 18.75 cm behind it; required, the curvature of the mirror. In this case q is negative because the image is known to be virtual. Then $1/75 - 1/18.75 = 2/r$. Hence $2/r = -1/25$, and $r = -50$, so the mirror is convex, as shown by the negative sign.

Since a convex mirror tends to reflect light in divergent rays, it can have no real principal focus for a parallel beam. However, a plane

wave means $p = \infty$, and then $q = -r/2$, so that a virtual image is formed between the center of curvature and the mirror. This is a virtual principal focus F at a distance f from the mirror, so that equation (3) of Article 413 applies to convex mirrors provided f is made negative in applying it to numerical calculations.

416. Construction of images. We have so far examined only the case of images and objects of points on the axis of a spherical mirror. But it is easy to extend this case to images of other points not too far from the axis, and so construct images of finite objects.

As was shown in Fig. 23, all the rays diverging from O are brought approximately to I by a concave spherical mirror of small aperture.

If O were then removed to infinity, all of its rays which reach the mirror, as shown in Fig. 25, would pass through the principal focus at F . Therefore any ray close to the axis and parallel to it passes through the principal focus after reflection, whether or not it came from an infinite distance. Now consider the object AB in Fig. 26. A ray l from the point A and parallel to the axis is reflected through F as indicated, and intersects the ray m . This latter passes through C and is therefore returned upon itself. So the point of intersection A' is located, and this is the image of A .

In the same manner, the image of B may be located at B' , and so for any other points of the object. Thus we find that this real image is inverted and smaller than the object. It is also perverted if it extends into space at right angles to the plane of the diagram. If the object had been $A'B'$, then the image would be AB , and larger than the object, but also inverted. If AB intersects the axis at C , then the

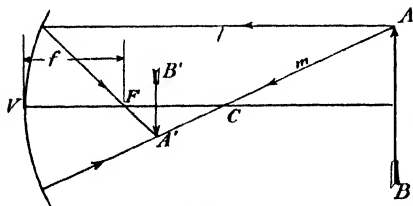


Fig. 26

image lies there also and is the same size. Therefore objects outside C form smaller images between C and F . Objects between C and F form larger images outside of C .

Objects between F and the mirror form virtual images. These may be constructed as in Fig. 27. Here the ray l parallel to the axis passes through F after reflection, and when produced behind the

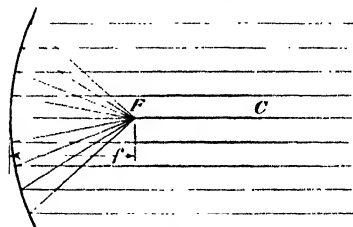


Fig. 25.

mirror, determines A' by its intersection with m , which passes through C . Then the eye in front of the mirror sees an apparent object at A'

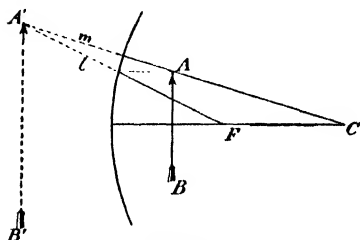


Fig. 27.

because it cannot distinguish between reflected and original rays, and all the reflected rays which started from A seem to diverge from A' . The point B' is found in the same manner, and the constructed image is seen to be erect and larger than the object.

If the mirror is convex, the construction depends upon the location of the virtual principal focus. The same two rays are again used. The solid lines in Fig. 28 show their actual direction before and after reflection, and the dotted lines are the reflected rays produced backward to their point of intersection. As in the preceding case, the eye sees A as at A' , because the rays seem to originate there. But now the image is nearer the mirror than the object, as was shown to be true in the preceding article, and it is obviously smaller.

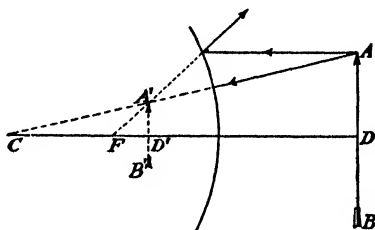


Fig. 28.

417. Relative size of image and object. In all the constructions described above, the point of the object lying on the axis, as D in

Fig. 29, has an image point also on the axis at D' . Also, in each case one ray or its production passes undeviated through the three points A , A' , and C ; that is, object point, image point, and center of curvature.

In Fig. 29, the ray AO from the object point A is reflected at O to A' making equal angles i and r with the normal OCD . Therefore the triangles AOD and $A'OD'$ are similar, and $AD/A'D' = OD/OD'$. The triangles ACD and $A'CD'$ are also similar, and $AD/A'D' =$

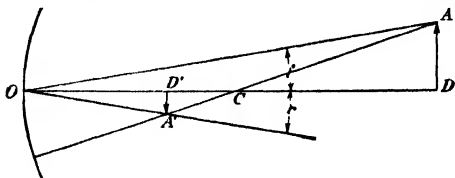


Fig. 29.

CD/CD' . But OD is the object distance p , and OD' is the image distance q . Then, setting u and v equal to the object and image distances from C respectively, we obtain the ratio of the sizes (indicated

by L_o and L_i), of object to image given by the two-fold relation

$$\frac{L_o}{L_i} = \frac{p}{q} = \frac{u}{v}.$$

In a similar manner the same relations are found to be true when *virtual* images are formed in either concave or convex mirrors, although in these cases the ratios $p:q$ and $u:v$ are negative, since the image lies behind the mirror. We may then make the general statement that the magnitudes of corresponding lengths of object and image are to each other as their distances from the center of curvature, or from the vertex of the mirror when measured along its axis.

418. Spherical aberration. In the preceding discussion it was assumed that the mirror was one of such small angular aperture that the approximation used in deriving the mirror formula was justifiable. But if the mirror has a large aperture, this is no longer the case, and rays parallel to the axis do not in general pass through the principal focus, but cut the axis between F and the mirror. This is known as **spherical aberration**, and the effect is greater the farther the ray is from the axis.

In Fig. 30 is shown a mirror of large aperture with parallel rays whose reflections cross the axis nearer and nearer the mirror as they recede from the axis. The mutual intersections of these rays form two luminous curves known as **caustic curves**, which meet in a cusp at F . A familiar example is the reflection from the inner surface of a cup nearly full of milk, when the light shines almost horizontally across the edge and is reflected by the opposite concave surface down onto the milk.

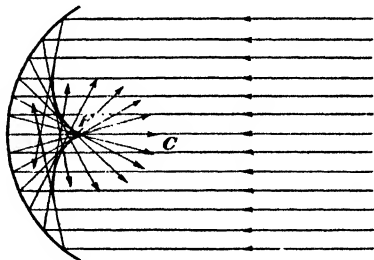


Fig. 30.

There it forms very striking caustics with a well-defined cusp.

419. Parabolic mirrors. Spherical aberration may be entirely eliminated for objects at a great distance by using a mirror whose surface is not spherical but formed instead by rotating a parabola about its own axis. The parabolic curve in Fig. 31 is the section of such a mirror. It is a well-known property of the parabola that the normal to the curve at any point bisects the angle between a line drawn to that point parallel to its axis, and a line drawn through the focus of the parabola. Therefore the angles β and γ are equal re-

ardless of the distance of the line Ad from the axis. But this fulfills the law of reflection; therefore all rays parallel to the axis pass through F , and conversely, all rays emanating from F are parallel to the axis after reflection.

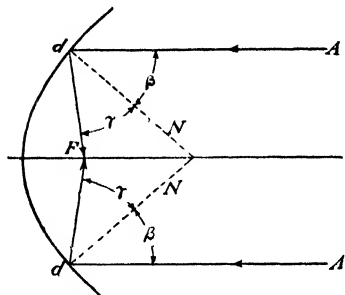


Fig. 31.

The parabolic reflector may therefore be used either to concentrate a plane wave (parallel rays) from a source a long way off at a sharply defined focus, or to produce a cylindrical beam from a point source at F . This latter is much the more important function of the mirror;

such reflectors are used in automobile headlights and in searchlights of all sorts. Generally, a slightly divergent beam is needed, instead of a strictly cylindrical one, and this may be produced by placing the light (usually an electric arc for searchlights) at a point slightly inside or outside F . If it is inside, the beam diverges at once; if it is outside, it is brought to a focus at some point P beyond the mirror and then diverges indefinitely as shown in Fig. 32, where S is the source. By bringing S nearer and nearer to F , P steadily recedes, and the angle α of divergence is steadily diminished.

The other use of a parabolic reflector, in concentrating a plane wave on a point focus, is illustrated in the great reflecting astronomical telescopes. These are equipped with mirrors having a parabolic figure rather than a spherical one, to eliminate spherical aberration.

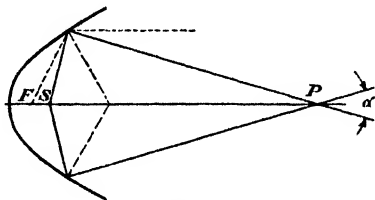


Fig. 32.

SUPPLEMENTARY READING

- J. Valasek, *Elements of Optics* (Chap. 5), McGraw-Hill, 1928.
 J. P. C. Southall, *Mirrors, Prisms and Lenses* (Chap. 2), Macmillan, 1933.
 R. W. Wood, *Physical Optics* (Chap. 2), Macmillan, 1934.
 T. Preston, *The Theory of Light* (Chap. 4), Fifth Edition, Macmillan, 1928.

PROBLEMS

1. A concave spherical mirror is 76 cm from an object, and forms a real image on a screen 133 cm distant. What is the mirror's radius of curvature?
Ans. 96.7 cm.

2. A luminous source is placed on the axis of a concave spherical mirror whose radius of curvature is 80 cm. The object distance is 30 cm. What kind of image is produced, and how far is it from the mirror? *Ans.* Virtual; 120 cm.

3. If the mirror in Problem 2 is convex, where is the image? *Ans.* 17.1 cm behind the mirror.

4. If the object is one centimeter long in Problems 2 and 3, how long are the images? *Ans.* 4 cm; 0.57 cm.

5. A concave spherical mirror has a radius of 120 cm. Where should an object be placed to form a real image having half its linear dimensions? *Ans.* 180 cm from the mirror.

6. Where should the object be placed with reference to a concave mirror of 96 cm radius to form a virtual image whose linear dimensions are eight times as large? *Ans.* 42 cm from the mirror.

7. It is desired to focus the image of an arc light formed by a concave spherical mirror on a screen 16 ft. from the arc. If the mirror has a radius of 4 ft., where should it be placed? *Ans.* 2 ft. 3 in. from the arc.

8. The sun subtends an angle of approximately 32 minutes of arc. What is the diameter of its image formed by a concave spherical mirror whose radius is 240 cm? *Ans.* 11.16 mm.

9. The radius of curvature of a convex mirror is 16 in. Where is the image of an object 4 ft. from the mirror? If the object is 6 in. long, how long is the image? *Ans.* 6.86 in. back of the mirror; 0.86 in.

10. A convex mirror of 16 in. radius forms an image of a landscape. What is the ratio of the angle which distant objects subtend at the eye, compared to the angle subtended by their images when the eye is 10 in. from the mirror? *Ans.* 9:4.

CHAPTER 32

Refraction at a Plane Surface

420. Snell's law. The Dutch astronomer Willebrod Snell (1591–1626) was the first to discover the law of refraction, proved by means of Huygens' construction in Article 329. Snell did not publish the results of his investigation, and they were not known until after his death. In the meantime, Descartes had made the same discovery and announced the law that *if the refracted ray and the incident ray continued through the point of incidence be intercepted by any line parallel to the normal to the surface at the point of incidence, the length of the intercepted portion of the refracted ray is in constant ratio to the length of the intercepted portion of the incident ray.* This is shown in Fig. 33, where db and pc are the incident and refracted rays, and their intercepts, by the line abc drawn parallel to the normal NN , are pb and pc respectively. In accordance with Descartes' statement, the ratio $pc:pb$ is constant. But $pc \sin \beta = ap = pb \sin \alpha$; therefore $pc/pb =$

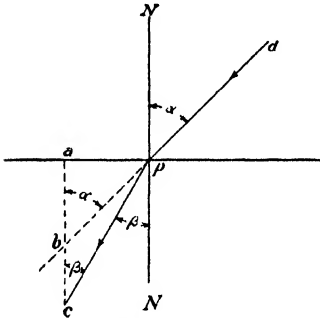


Fig. 33.

$\sin \alpha / \sin \beta = n$. This constant n is known as the **index of refraction**, and is equal to the ratio of the sine of the angle of incidence to the sine of the angle of refraction.

421. Causes of refraction. The proof in Article 329, based on wave motion, shows that the law of refraction is due to the change of velocity when waves enter a different medium across an interface. If the waves move more slowly after crossing the bounding surface between two media, their direction of propagation (ray) is bent toward the normal to the surface. If they move faster, the ray is bent away from the normal. This conclusion has been borne out by numerous experiments of all sorts, but it is worth noting that the corpuscular theory held by Newton could also account for the bending of light in accordance with Snell's law. In order to do so, it was necessary to

assume that the corpuscles which constituted light were attracted more by denser media than by rarer, and that in consequence, their *velocity increased* when they entered the denser medium.

A crucial experiment that proved Newton's hypothesis to be wrong was performed by Foucault in connection with his determination of the velocity of light, described in Article 399. He inserted a tube filled with water between the fixed and rotating mirrors. This tube was closed at the ends with glass plates, so that the beam of light passed through it from end to end, both going out and coming back. With the tube in position, the displacement of the beam, as viewed in the telescope, was greater than when the path was only through air. This proved conclusively that dense media refract light by decreasing its velocity, an assumption needed in explaining refraction by means of the wave theory. Therefore, even those who were still unconvinced by Fresnel's experiments, were obliged at last to abandon the corpuscular hypothesis.

422. Refraction by a parallel-sided plate. If a ray of light meets a parallel-sided transparent plate at right angles, it passes through undeviated, as would be expected from Snell's law, but if it is inclined at some other angle, it is bent on entering the medium, and then bent again just as much the other way on leaving it, so that its final and original directions are parallel to each other. This is shown in Fig. 34, where a ray of light, inclined at an angle α to the normal, enters and passes through a slab of glass. The angle of refraction, after passing through the surface AA , is β , which equals β' (opposite interior angles), the angle of incidence at the second surface. Finally it emerges at an angle α' . But $\sin \alpha / \sin \beta = n$, and $\sin \beta' / \sin \alpha' = n'$, where n' is the reciprocal of n , since the ratio of the velocities is reversed in passing from dense to rare. Therefore $\sin \alpha / \sin \beta = \sin \alpha' / \sin \beta'$, and since $\beta = \beta'$, $\alpha = \alpha'$, so that the emergent ray is parallel to the incident ray. But it has been displaced sideways through a distance which depends upon n , α , and t , the thickness of the slab. Let d equal the sidewise displacement pa , and let q equal the length of the path pb ; then $d = q \sin (\alpha' - \beta')$, and $q = t / \cos \beta'$; therefore $d = t \sin (\alpha' - \beta') / \cos \beta'$, where β' is known if n and α' are given.

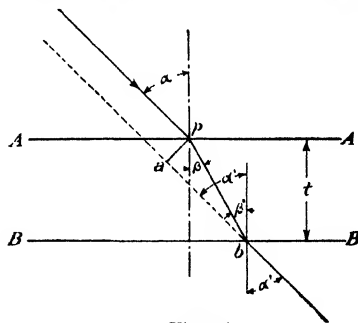


Fig. 34.

423. Total reflection. When light passes from an optically denser into an optically rarer medium, it is partly reflected at the interface, as shown in Fig. 35, and partly refracted at an angle β that is greater than α . The amount of light reflected internally increases as α increases until, when β exceeds 90° , none of it emerges, and the reflection becomes *total*. That is, all the energy of the original beam not absorbed by the medium comes back into it again. The limiting case, when $\beta = 90^\circ$, corresponds to a certain critical value of α known as

the **critical angle**. This differs with different media, and its value may be calculated from the relation $n = \sin \beta / \sin \alpha$, where the larger angle is in the numerator, as usual. When α has reached its critical value, $\beta = 90^\circ$, and $\sin \beta = 1$; therefore $\sin \alpha_c = 1/n$, where α_c is the critical angle. Thus in a kind of glass whose refractive index is 1.5, $\sin \alpha_c = 0.666+$, and $\alpha_c = 41^\circ 48'$.

This kind of reflection is much more satisfactory than that from ordinary mirrors, both because there is less absorption of light, and because ordinary mirrors give two images, one due to reflection at the silvered back, and one (much fainter) at the outer surface of the glass. Total reflection, on the other hand, occurs at only one surface, and there is only one image. If a prism is cut as shown in Fig. 36, and if its critical angle is less than 45° , then the incident beam may be bent through 90° without sensible loss of intensity, and the image is not confused by a second fainter one. Such prisms are used in prism binoculars and in certain telescopes where ordinary mirrors or polished metal surfaces would be most unsatisfactory.

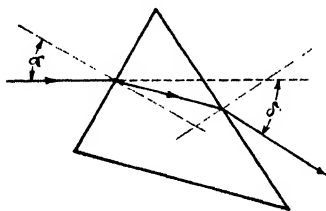


Fig. 37.

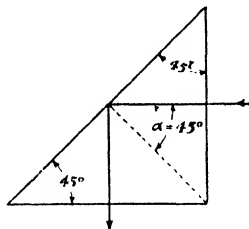


Fig. 36.

424. Refraction by a prism. If a ray of light enters a glass prism at an angle of incidence α , as shown in Fig. 37, it is bent toward the normal to the refracting surface, passes through the prism, and is bent away from the normal on emergence, undergoing a total devia-

tion δ . The maximum possible deviation is obtained when the emergent ray just grazes the second surface; which means that δ increases as α decreases, and is maximum when α is minimum. Now as optical paths are reversible, we might consider the emergent ray as the incident ray, so that the angular deviation is a maximum when the angle of incidence is also a maximum, or 90° . Between these two maxima must be a minimum value, and this is found when the incident and emergent rays make equal angles with their respective surfaces, and the path through the prism is perpendicular to the bisector of the angle A made by the intersection of the two faces considered. This arrangement is shown in Fig. 38, where the angles of entry and emergence are equal, and the angles β are also equal to each other and

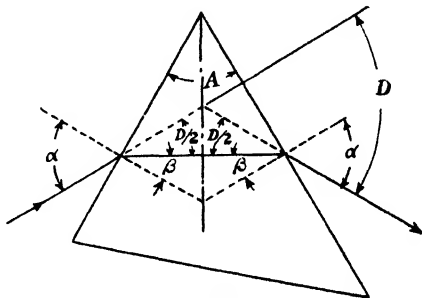


Fig. 38.

to half the refracting angle A of the prism, because their sides are mutually perpendicular to those of $A/2$. Consideration of the symmetry of the diagram shows that the incident and emergent rays, when produced, meet at a point on the bisector of the angle A ; therefore they make equal angles with the refracted ray between them, and if the total deviation is D , the deviation at each interface is $D/2$. Then, as they are vertical angles, $\alpha = \beta + D/2$, and substituting $\beta = A/2$, we have

$$\alpha = \frac{A}{2} + \frac{D}{2} = \frac{1}{2}(A + D).$$

But

$$n = \frac{\sin \alpha}{\sin \beta}.$$

$$\therefore n = \frac{\sin \frac{1}{2}(A + D)}{\sin \frac{1}{2}A}. \quad (1)$$

This equation makes it possible to calculate the index of refraction in terms of two angles, which are easily measured. It should be especially noted, however, that this equation is true only for the symmetrical case supposed, when the angle of deviation is a minimum. Therefore, in finding n by this method, the prism is turned back and forth until it is evident that the light is bent less than at any other setting of the prism with respect to the incident beam.

If the refracting angle of the prism is very small, the deviation is small also; therefore in such cases we may take the angle for its sine, and then

$$n = \frac{A + D}{A}, \quad (2)$$

which greatly simplifies the calculation.

SUPPLEMENTARY READING

- R. A. Houstoun, *Intermediate Light* (Chap. 3), Longmans, Green, 1925.
 T. Preston, *The Theory of Light* (pp. 95-107, 129-134), Fifth Edition, Macmillan, 1928.
 J. P. C. Southall, *Mirrors, Prisms and Lenses* (Chapters 3, 4), Macmillan, 1933.

PROBLEMS

1. It is just possible to see the bottom of a square jar filled to the brim with water, when viewed at an angle of 30° from the vertical. The jar is 8 in. deep and 3.2 in. square. What is the observed index of refraction of the water? *Ans.* 1.35.
2. The critical angle of a slab of glass is found by experiment to be 44° . What is its refractive index? *Ans.* 1.44.
3. A small incandescent lamp is immersed to a depth of 64 cm in a tank of water. What is the diameter of the circle at the surface of the water which bounds the emergent cone of light? (Take $n = 1.33$.) *Ans.* 146 cm.
4. A prism whose refracting angle is 60° causes a minimum deviation of 46.3 in a monochromatic beam of light. What is its refractive index? *Ans.* 1.6.
5. What should be the refracting angle of a prism whose index is 1.6, in order to have a minimum deviation of 10° ? *Ans.* 16.24 by equation 1, Article 424; 16.7 by equation 2.
- *6. A ray of light passes through a parallel-sided slab of glass 4 cm thick. The angle of incidence is 45° , and $n = 1.5$. Calculate the sideways displacement of the emergent ray. *Ans.* 13 mm.

CHAPTER 33

Lenses

425. Types of lenses. Lenses are usually made of glass, but any transparent medium having an index of refraction greater than air will do; for special purposes lenses are sometimes made of quartz or rock salt. They are bounded by curved surfaces which are usually spherical, but may be cylindrical or both combined. We are familiar with lenses in eyeglasses, magnifying glasses, photographic cameras, and so forth. Their purpose is to change the curvature of the wave front of the light which falls upon them. In the eyeglass, this adapts the curvature to an abnormal eye. In the magnifying glass or simple microscope, the change of curvature makes an object look larger. In a camera, the lens "focuses" an object upon a sensitive plate, where an image is formed. There are six possible types of lens defined by spherical or plane surfaces. Sections of these types are shown in

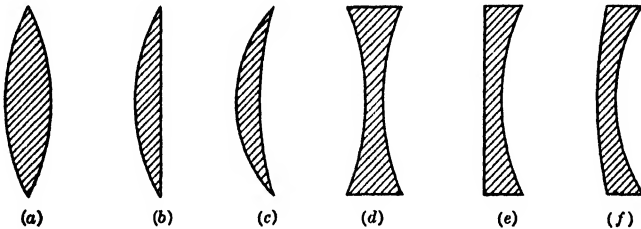


Fig. 39.

Fig. 39, and are known as (a), double convex; (b), plano-convex; (c), meniscus; (d), double concave; (e), plano-concave; and (f), concavo-convex. The first three are thicker at the center than at the edge, and as we shall see, are converging lenses, that is, tending to make divergent rays converge. The others (d, e, and f) are thinner at the center and are diverging, tending to make divergent rays diverge still more. In terms of wave-front curvature, a converging lens tends to decrease or reverse the curvature of the wave front from a real source, while a diverging lens tends to increase it.

426. Object and image relations. The equation to be used in locating the image (or object, if the image distance is given) is the same as that for mirrors, namely,

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f}. \quad (1)$$

For converging lenses, commonly called positive lenses, the sign of f is positive, and for diverging lenses, commonly called negative lenses, the sign of f is negative. When q , the image distance, is positive, the image is real. When it is negative, it is virtual. Positive lenses are most commonly used, and the relations between image and object present six cases of especial interest, similar to those of a concave mirror. These are:

(a) A telescope forms a real image of a distant star. Here p is practically infinite; therefore $1/p = 0$ and from equation (1), $q = f$. The image is then at the principal focus.

(b) A photographic camera forms a real image of objects at a finite distance, and usually at a distance from the lens greater than $2f$. Then $p = 2f + e$, where e is some finite quantity, and

$$\frac{1}{q} = \frac{1}{f} - \frac{1}{2f + e}, \quad (2)$$

or
$$q = 2f - \frac{fe}{f + e}. \quad (2')$$

These show that the image lies between f and $2f$.

(c) In making photostatic copies of manuscripts, the image and object have the same size. This occurs when $p = 2f$. Then $e = 0$ in equation (2'), and $q = 2f$.

(d) A projection lantern forms a distant real image of the nearby lantern slide, which must lie between f and $2f$. This case is the reciprocal of case (b). As object and image are always interchangeable in geometrical optics, the image lies between $2f$ and infinity.

(e) A flashlight, adjusted to send out parallel rays, forms an image of the glowing filament at infinity. This is the reciprocal of case (a), q is infinite, and $p = f$. Hence an object at the principal focus forms a real image at an infinite distance.

(f) A magnifying glass forms an enlarged image on the same side of the lens as the object when the object lies between the principal focus and the lens. Solving (1) for q , we have

$$q = \frac{pf}{p - f}, \quad (3)$$

and when p is less than f , q is negative and virtual.

These six cases are illustrated in Fig. 40. They may be summarized as follows: As the object moves from infinity toward a converging lens, the image, starting at a distance f on the other side, moves farther away until it reaches $2f$, when the object is at $2f$ also. The

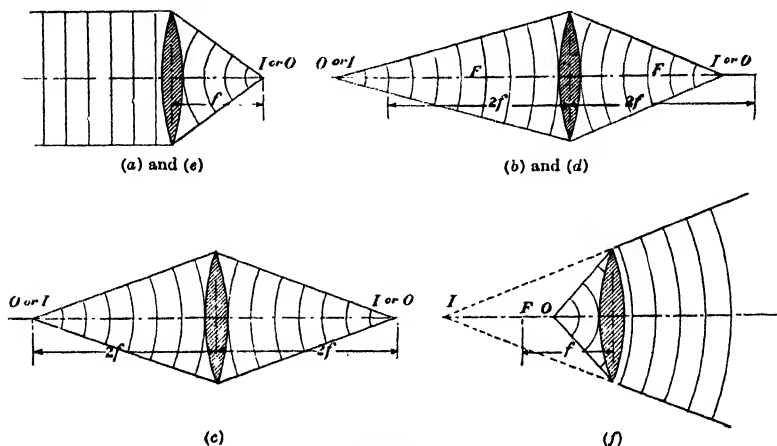


Fig. 40.

object and *real* image are then $4f$ apart, which is their nearest possible approach. If the object is moved still nearer to the lens, the image moves steadily outward, until with the object at the principal focus, the image is at infinity. After that the image is virtual, with the object inside of the principal focus.

In the case of a diverging lens, f is negative, and if the object is real, q is necessarily negative also. That is, diverging lenses can form only virtual images of real objects. If the object is at infinity, the

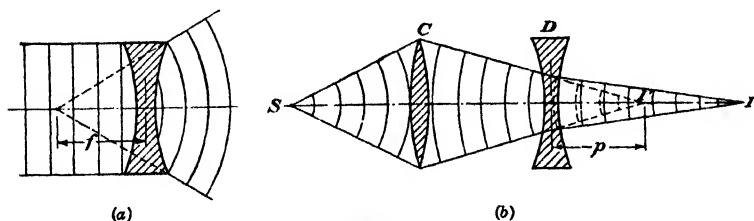


Fig. 41.

image is at the principal focus, which is on the same side of the lens as the object, and therefore may be regarded as a virtual focus, as shown in Fig. 41 (a).

A diverging lens can, however, produce a real image when in the path of a sufficiently convergent beam. Its diverging power is then insufficient to overcome the convergence of the incident light. This arrangement is shown in Fig. 41 (b), where the apex of the dotted cone is the real image, I' , of the source S formed by the lens C acting alone. The point I is the real image formed at a greater distance by the combination, where I' is the virtual object at a distance $-p$ from lens D . The location of I is then found by solving the usual equation for q , with f and p both negative.

427. Problems concerning simple lenses. The lens equation contains three quantities, each of which may be calculated if the other two are given. For a given lens the focal length is supposed known, and we may then find the image distance for a given object distance, as in the case just cited. But it is also possible to find the focal length of a lens by observing q , with the object a given distance from the lens. Or we may find where the object must be in order to form an image in a given position, when f is known.

A fourth quantity that may be given or required is the magnification. This, as with mirrors, is the ratio $q : p$.

As an illustration of the most usual problem, suppose a positive lens, whose focal length is 12 cm, is placed 15 cm from a lighted candle. Required, the image distance. Then

$$\frac{1}{15} + \frac{1}{q} = \frac{1}{12},$$

$$q = \frac{12 \times 15}{15 - 12} = 60 \text{ cm},$$

and the magnification is $60:15 = 4$.

If the lens had been diverging, then $f = -12$, and

$$\frac{1}{15} + \frac{1}{q} = -\frac{1}{12}.$$

$$\therefore q = \frac{12 \times 15}{-15 - 12} = -6\frac{2}{3} \text{ cm},$$

where the negative sign indicates a virtual image. If q is given and p required, we must know whether the image is real or virtual. Or, what is the same thing, we must know whether it is inverted or erect, because converging lenses, like concave mirrors, form inverted real images and erect virtual images, while diverging lenses, like convex mirrors, form only erect images.

Let it be required to find the object distance when a positive lens of 10 cm focal length forms an erect image at 15 cm. As the image is erect, it must be virtual; therefore

$$\frac{1}{p} - \frac{1}{15} = \frac{1}{10},$$

$$p = \frac{10 \times 15}{10 + 15} = 6 \text{ cm},$$

and the magnification is $15:6 = 2.5$.

But if the image had been real at 15 cm, then q is positive and

$$\frac{1}{p} + \frac{1}{15} = \frac{1}{10}.$$

$$\therefore p = \frac{10 \times 15}{15 - 10} = 30 \text{ cm}.$$

The magnification is 0.5, which is a reduction of size, as is to be expected, because q lies between f and $2f$, and the object is outside of $2f$, as we have seen.

428. Minimum distance between object and image. Since an object at infinity has an image at the principal focus F , while an object at F forms an image at infinity, and since, for all other positions of the object greater than f , the image is at a finite distance from the lens, it follows that there must be a minimum distance between object and image. From purely geometrical considerations of the reversibility of p and q , this minimum must be found when $p = q$. But when $p = 2f$, $q = 2f$; therefore the minimum distance is $4f$. This is proved algebraically by Preston† in the following ingenious manner: In the case of a real image formed by a converging lens, $1/q + 1/p = 1/f$. Squaring both sides of the equation, we obtain

$$\left(\frac{1}{p} + \frac{1}{q}\right)^2 = \left(\frac{1}{p} - \frac{1}{q}\right)^2 + \frac{4}{pq} = \frac{1}{f^2}.$$

But as f is constant for a given lens, $4/pq$ must be a maximum, or pq a minimum, when $p = q$, because this reduces the parenthesis to zero.

Also, since $\frac{1}{f} = \frac{p+q}{pq}$ is constant, $p+q$ is a minimum when pq is a minimum; therefore $p+q$ is a minimum when $p = q$, or when $p+q = 4f$.

† T. Preston, *The Theory of Light* (problem 1, p. 114), Fifth Edition, Macmillan, 1928.

This shows that it is impossible to form a real image of a real object on a screen, if the distance from object to screen is less than $4f$, which is a conclusion of considerable practical importance.

429. Experimental measurement of focal length. A rough method for finding the value of f of a convergent lens consists in forming the image of a distant object or of the sun itself upon a screen. The distance of the image from the screen is then approximately equal to the focal length. But this important constant of a lens may be found with much greater precision in the following manner: The lens is arranged to form a real image of a nearby luminous source on a screen, in such a way that the image is either larger or smaller than the object, as in Fig. 42. Then $p + q$ must be greater than $4f$, for if equal to $4f$,

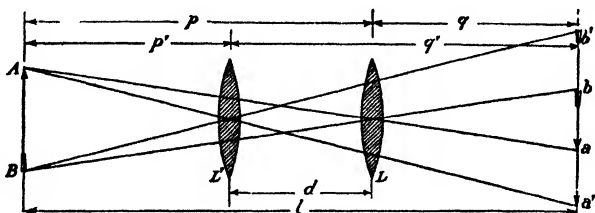


Fig. 42.

image and object would be of the same size. With object and screen fixed, the conjugate positions of the lens are readily found, one giving an image larger and the other smaller than the object. If the distances d and l are carefully measured, f may be calculated as follows: The equations, as usual, are $1/p + 1/q = 1/f$, and $1/p' + 1/q' = 1/f$ for the two positions of the lens. But as the image and object are interchangeable, $p = q'$ and $q = p'$, so that one equation answers for both cases. Also, $p + q = l$, and $p - q = d$. Solving these equations for p and q , we have

$$p = \frac{l + d}{2},$$

and

$$q = \frac{l - d}{2}.$$

Substituting these values in the lens equation, we obtain

$$\frac{2}{l + d} + \frac{2}{l - d} = \frac{1}{f},$$

and

$$\frac{4l}{l^2 - d^2} = \frac{1}{f}.$$

Thus f may be found in terms of l , which may be measured with precision, and of d , whose determination depends upon the skill with which the object is focused in the two positions.

430. The general formula for simple optical systems. Before we derive the lens formula and show how to calculate f from the known constants of the lens, it is well to obtain a more fundamental relation which applies to both mirrors and lenses, and from which the lens formula easily follows.

It is desired to find how the curvature of a wave front is altered when it passes from one medium into another of different optical density. We shall suppose that the bounding surface between the two media (the *interface*) is spherical, though it may be plane when the radius of the sphere is infinite. We shall also adopt a convention opposite to that used with mirrors, and regard distances measured from the interface as positive when they are measured away from the source, and distances measured toward the source as negative.

In Fig. 43 let the spherical surface ABC , of radius R , separate the medium numbered 1 (air) from the denser medium numbered 2. Let an object O be at the negative distance $-p$ from the pole B of the surface, and consider a ray OP which meets the surface at P , making the angle of incidence i with the normal DP produced. This ray

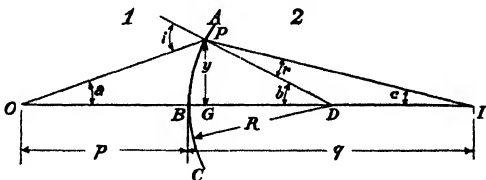


Fig. 43.

is then refracted toward DP , with which it makes the angle of refraction r , and at the image point I it makes an angle c with the axis. From P a line y is drawn normal to the axis and meeting it at G .

Let q represent the image distance BI , and let a and b represent the angles POB and PDB respectively. Then by referring to the triangle PDO , we find that the exterior angle i equals the sum of the opposite interior angles, or

$$\angle i = \angle a + \angle b. \quad (1)$$

Similarly in triangle PDI

$$\angle b = \angle r + \angle c,$$

or

$$\angle r = \angle b - \angle c. \quad (2)$$

If we assume that the angles i and r are small, as is generally the case

in such problems, we may set their sines equal to their angles measured in radians without serious error;

then
$$\frac{\sin i}{\sin r} = \frac{i}{r} = n.$$

Dividing (1) by (2), and substituting n , we obtain

$$\frac{a+b}{b-c} = n. \quad (3)$$

Taking the sines or tangents of small angles equal to the angles, also taking $OG = p$ and $GI = q$ approximately, then $\angle a = y/p$, $\angle b = y/R$, and $\angle c = y/q$, approximately. Substituting these values in (3), transposing, and canceling the common term y , we obtain

$$-\frac{1}{p} + \frac{1}{R} = n \left(\frac{1}{R} - \frac{1}{q} \right). \quad (4)$$

We shall now define a new quantity, ρ . This represents the ratio of the velocity of a beam of light *after* passing into a new medium, to the velocity *before* that change. In the case considered above, the second medium is denser than the first, the second velocity is therefore the lower one, and ρ is equal to the reciprocal of n , which always expresses the ratio of the higher to the lower velocity regardless of which came first.

We shall also introduce curvatures in place of the reciprocals of the radii of curvature as indicated below. Thus in the case represented by Fig. 43,

$\rho = \frac{1}{n}$, the reciprocal of the refractive index.

$C_1 = \frac{1}{p}$, the curvature of the wave front incident on ABC .

$C_2 = \frac{1}{q}$, the curvature of the wave front after passing through ABC .

$S = \frac{1}{R}$, the curvature of the bounding surface, or interface.

Equation (4) may now be written

$$-C_1 + S = \frac{1}{\rho} (S - C_2),$$

which reduces to the convenient form

$$C_2 = \rho C_1 + (1 - \rho)S. \quad (5)$$

This gives the curvature of the modified wave front in terms of known quantities, while the reciprocal of C_2 gives us the image distance q measured from the pole B .

Equation (5) applies to all cases of reflection and refraction which come within the scope of geometrical optics. For instance, let us consider reflection by a plane mirror whose curvature S is zero. In this case the change in velocity involves only a reversal of direction, but no change of speed, so $\rho = -1$. Therefore, setting $S = 0$, and $\rho = -1$ in equation (1), we obtain $C_2 = -C_1$, which means that the reflected and original wave fronts have equal and opposite curvatures, so that their centers (image and object) are equidistant from the interface and on opposite sides, as has already been proved by the ray construction. Again, if we consider a concave spherical mirror, ρ is still equal to -1 , but S is no longer zero and is negative because the mirror's center of curvature lies toward the source. Therefore $C_2 = -C_1 - 2S = -C_1 - 2/r$. Then set $C_1 = -1/p$, where p is object distance, and $C_2 = -1/q$, where q is the real image distance, both being negative because they are measured toward the source. Making these substitutions, we obtain

$$\frac{1}{p} + \frac{1}{q} = \frac{2}{r},$$

which was also proved by ray construction.

Another interesting application of the formula is to the case of an object on one face of a plane parallel-sided slab of a transparent medium, such as a layer of water when the object is seen through the medium. In this case $S = 0$, and $\rho = n$, the index of refraction of the substance, because we are now passing from a denser medium into air. Then $C_2 = nC_1$, or $q = p/n$, which means that the object appears nearer than it really is. In other words, the apparent distance below the surface is one n th of the true distance.

431. Proof of the general lens formula. In passing through a lens, a beam of light encounters two bounding surfaces instead of one, and the curvature of the wave front suffers two modifications. In Fig. 44

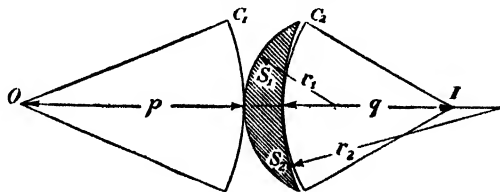


Fig. 44.

the original wave front starting at O has a curvature $-C_1$ when it meets the lens. After emerging from the lens, its curvature is C_2 , which is

here assumed positive, so that a real image is formed at I at the positive distance q from the lens. The curvatures of the two lens surfaces are S_1 and S_2 . Both have purposely been made positive with their centers of curvature away from the source of light.

The original beam, after passing through the first interface, is modified in accordance with equation (5) of Article 430. As C_1 is negative, this becomes

$$C' = -\rho_1 C_1 + (1 - \rho_1) S_1, \quad (1)$$

where ρ_1 is the ratio of velocities, in this case equal to $1/n$, and C' is the modified curvature within the lens. This modified wave front changes curvature in passing through the lens, but if the lens is thin in comparison with the image and object distances, the change may be neglected.†

We now apply the same equation to the passage of the beam through the second interface. This time C' takes the place of C_1 in equation (1), C_2 replaces C' , and ρ_2 replaces ρ_1 . Then

$$C_2 = \rho_2 C' + (1 - \rho_2) S_2. \quad (2)$$

Substituting the value of C' from (1) in (2), we obtain

$$C_2 = \rho_2 [-\rho_1 C_1 + (1 - \rho_1) S_1] + (1 - \rho_2) S_2. \quad (3)$$

Then substituting $1/n$ for ρ_1 , and n for ρ_2 , and collecting terms, we have

$$C_2 = -C_1 + (n - 1)(S_1 - S_2). \quad (4)$$

The quantity $(n - 1)(S_1 - S_2)$ or $(n - 1)(1/r_1 - 1/r_2)$ depends upon the index of refraction and the radii of curvature of the two surfaces. It is called the optical *power* of the lens, and may be represented by the Greek letter ϕ , so that (4) becomes

$$C_2 = -C_1 + \phi. \quad (5)$$

This shows that the lens may decrease the original curvature C_1 or reverse it, when ϕ is positive. It increases the original curvature when ϕ is negative. The power of a lens is measured by optometrists in **diopeters**. This is the numerical value of ϕ when the radii of the curvatures r_1 and r_2 , as well as the distances p and q , are measured in meters instead of centimeters. Converging lenses have positive optical powers, and tend to increase the curvature of a convergent

† As represented in Fig. 44, this would not be permissible. Here the lens is purposely made thicker than usual in order to make a clearer diagram having C_2 greater than S_2 .

wave front (plus before C_1), while diverging lenses have negative optical powers and tend to increase the curvature of a divergent wave front (minus before C_1 , as in equation (5)).

432. The simplified lens equation. If equation (5) is applied to a positive lens when the source of light is at infinity, the curvature C_1 of the incident wave is zero (plane wave front), and $C_2 = \phi$. This means that the modified wave front is convex toward the source (C_2 is positive) and its center lies on the opposite side of the lens, as shown in Fig. 40 (*a* and *e*). But we know that this center is the principal focus, so its distance f from the lens is the reciprocal of the curvature C_2 . Hence

$$C_2 = \frac{1}{f} = \phi,$$

or
$$\frac{1}{f} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right), \quad (1)$$

or the power of a lens is equal to the reciprocal of its focal length.

The incident wave front usually originates in a real "object," and its curvature is therefore convex toward the lens and negative. At the lens its radius becomes the object distance, and $C_1 = 1/p$. Similarly the emergent wave front has a center of curvature toward which it converges, or from which it diverges. The distance from the lens to this point is the image distance, and $C_2 = 1/q$. Then if we substitute these values for ϕ , C_1 , and C_2 , equation (5) of the preceding article reduces to

$$\frac{1}{q} + \frac{1}{p} = \frac{1}{f}, \quad (2)$$

as was stated without proof in Article 426. This equation assumes a real object at a negative distance $-p$ from the lens. But if we had taken a converging wave front having a virtual object point a positive distance p beyond the lens, all three variables would be positive and the equation would have the more general form $1/q - 1/p = 1/f$. The objection to this equation is that a real object has a negative object distance, and we have to remember to give p a negative sign, in nearly all lens problems, while in using (2) this is already taken care of.

If the lens is a diverging one, ϕ and its reciprocal f are both negative and the equation becomes

$$\frac{1}{p} + \frac{1}{q} = -\frac{1}{f}. \quad (3)$$

The changes in the curvature of the wave front, when a positive lens forms a real image, are shown in Fig. 45. The wave entering the lens is progressively retarded, and thus its curvature is reduced.

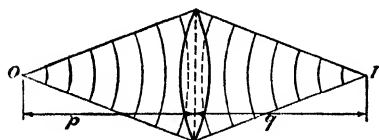


Fig. 45.

As it emerges, the portions which leave the retarding medium first speed up first, and so reverse the original curvature. If the lens is diverging or negative, the first face, if concave, may have very little effect upon the curvature

of the wave front. But when it emerges from the second face, the portions near the axis come out first, and speed up in air before the portions far from the axis. This results in increasing the original curvature, as shown in Fig. 46, and a virtual image is formed at I , the center of curvature of the emergent wave front.

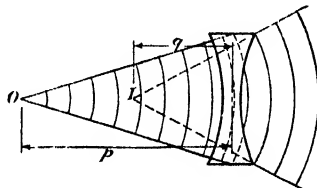


Fig. 46.

433. Optical center. A ray of light passing along the axis of a lens is of course undeviated, but there are other paths through a lens which involve no deviation, but only a slight lateral displacement, as when a ray goes through a parallel-sided slab. The intersection of these rays is known as the **optical center** of the lens, and is shown at P in Fig. 47, where the axial ray R intersects another drawn through the points A and B . These points are defined by the fact that two planes tangent to the lens surfaces at A and B are parallel to each other, as indicated in the sectional view given. Two planes fulfilling such a condition may be found for any kind of lens, but P is at the center only when the lens is symmetrical about that point, as in types (a) and (d) of Fig. 39, with both curvatures equal. Any ray passing through P is undeviated; also all undeviated rays must pass through this point.

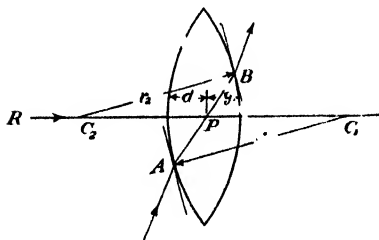


Fig. 47.

It is not very difficult to prove that the location of P is given by

$$g = \frac{r_2 t}{r_2 - r_1},$$

where g is as shown in Fig. 47, r_1 and r_2 are the radii of curvature of the faces of the lens, and t is its thickness measured along the axis. If the lens is plano-convex with the plane side facing the source, r_1 is infinite and $g = 0$. This means that P lies in the convex surface. Considering a meniscus lens as shown in Fig. 39, both curvatures are positive and r_2 is greater than r_1 . Therefore g is positive and greater than t , and P lies outside the convex face. In the case of a concavo-convex lens, both curvatures have the same sign as for a meniscus lens, and both may be taken as positive, but now r_1 is greater than r_2 ; therefore g is negative, which means that P lies outside the concave face.

434. Construction of images. We have so far discussed only point sources of light lying on the axis of the lens, but as in the case of mirrors, it is easy to construct an image of a real object point by point, if we make use of the intersection of two rays whose paths are known.

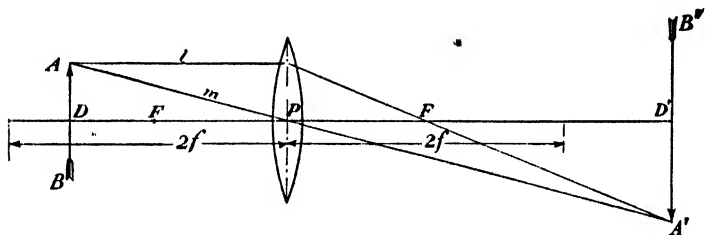


Fig. 48.

This is shown in Fig. 48, where a point such as A lies on what is called the secondary axis APA' , and its image must lie somewhere on this axis. Its exact position is found from the intersection of the two rays l and m . Thus l is drawn parallel to the axis, and after emergence it passes through F like all parallel rays when the lens has a small aperture. Ray m is drawn through the optical center, forming a secondary axis. It is undeviated, and its sidewise displacement is ignored because the lens is supposed thin.

In a similar manner we may locate the image B' of the point B , and so construct the entire image, which is inverted like all real images, and larger than the object. It is also perverted if it extends into space at right angles to the plane of the diagram. In the case shown, the object is between f and $2f$, so that the image is outside $2f$. But image and object may always be interchanged, so that $A'B'$ may be thought of as the object and AB its image, which is then smaller than the object.

If the object intersects the axis of the lens at F , it is said to lie in the focal plane of the lens. Then the rays from any point of the object are parallel after passing through the lens, and in general, inclined to its primary axis DPD' . Their direction is that of a secondary axis APA' , as already explained, and the image is infinitely distant, as shown in Fig. 49.

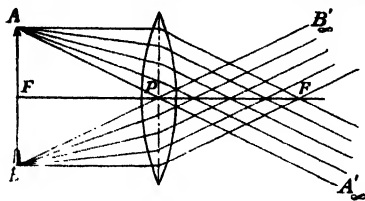


Fig. 49.

When the object AB is inside F , as in Fig. 50, then the construction of the virtual image calls for a backward production of the emergent rays, which diverge because the lens has not sufficient "power" to converge those originating so near it. To an eye situated to the right of the lens in Fig. 50, the rays shown appear to originate at A' instead of A , and the complete image $A'B'$ is seen at a greater distance than the object. It is enlarged and erect instead of being inverted like a real image.

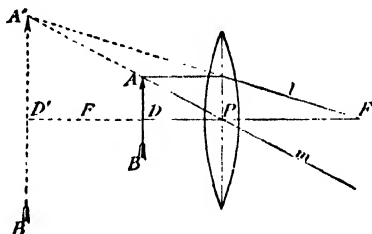


Fig. 50.

Images of real objects, formed by diverging lenses, are always virtual, and the principal focus is in a sense virtual also. Thus a ray drawn from A in Fig. 51, parallel to the axis, is bent away from it after emergence, but when produced backward it intersects the axis at F , and crosses the undeviated ray through P at A' , which is therefore the image of A . The complete image $A'B'$ is therefore nearer the lens than the object AB , and is erect but reduced in size.

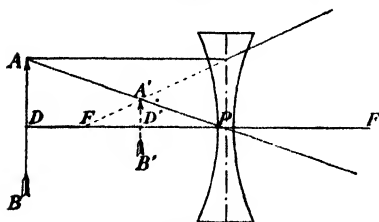


Fig. 51.

435. Magnification by lenses.

In all the preceding constructions we can find two similar triangles whose altitudes are similar portions of object and image. These in Figs. 48, 50, and 51 are ADP and $A'D'P$. Therefore $AD/A'D' = DP/D'P = p/q$, where p and q are the object

and image distances respectively. Since AD and $A'D'$ are corresponding parts of the object and image (half its length as illustrated), the ratio of the size of the image to that of the object is as

q is to p , or equal to the ratio of their respective distances from the optical center. An image nearer the lens than the object is therefore smaller, but if more distant, it is larger than the object. This ratio of image to object is called the magnification of the lens, but this purely linear relation must not be confused with another use of the word magnification as applied to perception by the eye, when the *apparent* size of both image and object are considered, as will be explained later.

436. Spherical aberration. A spherical wave front, produced by a point source on the axis of a converging lens, is no longer spherical when it emerges from the lens, a fact that is increasingly apparent as the aperture of the lens is taken larger and larger.

The emergent wave front is as shown in Fig. 52. The normals to this surface intersect in caustic curves of which it is the evolute. These caustics form a cusp at F , which is the principal focus of the lens, and a point where an ideal pencil of rays at the axis would be concentrated.

An important consequence of spherical aberration is the increasing *distortion* of the image of an object at increasing distances from the axis. This may be corrected by the

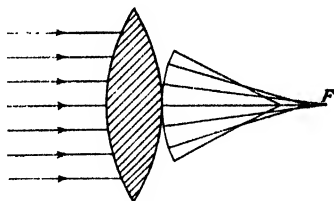


Fig. 52.

use of a “diaphragm” of small aperture placed close to the lens to shut off all the light except near the axis and prevent distortion at the expense of most of the illumination.

If such a diaphragm is moved from the lens toward the object, the image becomes brighter, but the marginal rays (far from the axis) now pass through. This produces “barrel distortion” of the image, as shown in Fig. 53 (a), where the object AB is a rectangular mesh seen edgewise. The diaphragm D shuts out those rays which would form the undistorted rectangle of altitude ab , but it passes the marginal rays which experience spherical aberration. These marginal rays are bent too much, as was shown in Fig. 52, and produce the images a' and b' . Thus the dimensions of the rectangle are reduced, and increasingly so at increasing distances from the center O . The point p , for example, approaches O through the distance pp' , while q , which is farther off, approaches O through the greater distance qq' .

If the diaphragm is on the other side of the lens, as in Fig. 53 (b), the image of the rectangle AB is stretched out by the marginal rays

to an altitude $a'b'$ instead of ab which would be the image's undistorted altitude. The result is "pincushion distortion" due to the stretching out of such points as p to p' , while q , more distant from the center, is stretched through a greater distance to q' .

If two similar lenses are used, and the diaphragm is placed midway between them, the two types of distortion tend to correct each other. This is usual in the compound lens of a photographic camera.

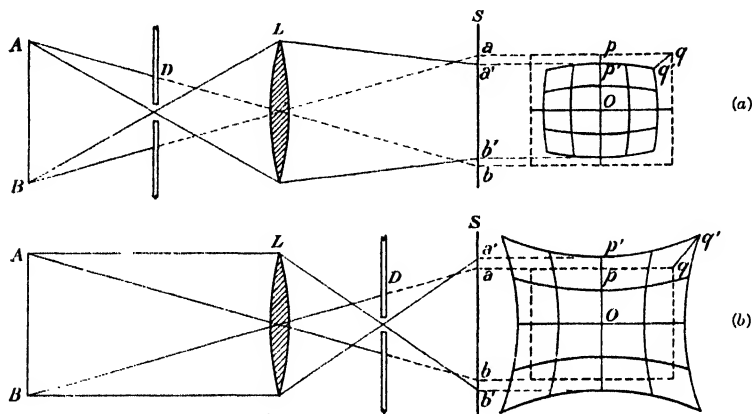


Fig. 53.

The two kinds of distortion described above may be strikingly demonstrated by placing the mesh in a converging beam of light whose focus is at the opening in the diaphragm. The diaphragm is then no longer needed, but the lens L must be placed with reference to the focus of the converging beam as if the focus were the opening in the diaphragm.

437. The correction of defects. For lenses of fairly large aperture, this is possible only when two or more lenses are combined. The simplest combination, as explained in Article 436, is a pair of similar positive lenses with a diaphragm midway between them. This corrects distortion, and if the diaphragm is correctly chosen, gives uniform illumination of the field. A very costly method is the grinding of lenses with nonspherical surfaces. In this way it is possible to eliminate spherical aberration completely for a specified object distance.

A very simple device, which results in a partial correction of aberration of rays parallel to the axis, is that of a plano-convex lens with its curved surface toward the source of light. A ray parallel to the

axis, which meets the lens near its outer edge, passes through the glass at an angle which may be perpendicular to the median of the angle at A , as shown in Fig. 54. If we regard the shaded portion of the lens as a small prism, such a ray experiences minimum deviation and intersects the axis at a point f very near the principal focus.

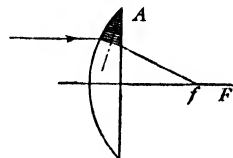


Fig. 54.

438. Thick lenses. When a lens is too thick to permit use of the approximation used in deriving the lens formula, it is still possible to calculate the position of an image, or the focal length, by a method devised by Gauss. This amounts to eliminating the central portion of the lens by means of the construction of the **principal planes**, ab and $a'b'$, in Fig. 55. From these the object and image distances, as well as the focal lengths, are measured, instead of from a plane through the optical center of the lens. The points H and H' , where the principal planes cut the axis, are called the **principal points** of the lens, and the object and image planes, cutting the axis at D and D' , are called **conjugate planes**, because every point in the D plane corresponds to an image point in the D' plane, if we neglect spherical aberration or other lens defects.

Now consider a ray from A which passes through the principal focus F of object space. This ray is parallel to the axis after emergence, and $aH = a'H'$. The ratio of these distances is obviously

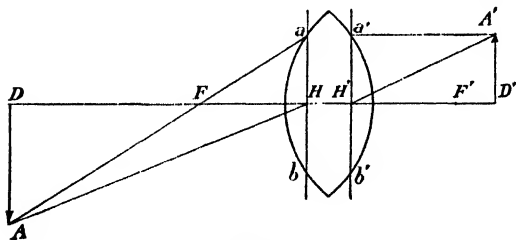


Fig. 55.

unity for this particular ray; therefore the principal planes are often called **unit planes**, because the conjugate points a and a' are equidistant from the axis. This is true of all such points; therefore H'

is the conjugate of H , and the rays AH and $H'A'$ are conjugate rays.

In double convex lenses, the principal planes are always inside the lens; in others they may lie on the surface or outside. Their approximate location for the different kinds of lens is shown in Fig. 56. If the positions of the planes are known, we may use the lens formula to locate an image point, by measuring p and q from the principal planes, and we may construct images as indicated in Fig. 55, using the rays

AH and $H'A'$, which are parallel but not in the same straight line, like the rays through the optical center of thin lenses.

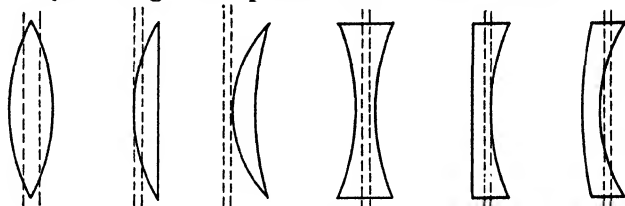


Fig. 56.

439. Image of a pencil of parallel rays inclined to the axis. Light from objects at a great distance has a plane wave front, and if the source is not a point, there are many such wave fronts inclined to each other. In other words, there are many bundles of parallel rays, each originating in a point of the object. Let l and m in Fig. 57 be two rays from a point of a distant object, and let l pass through the principal focus in object space, as shown. This ray emerges parallel to the axis so that $y = y'$. But the other ray, m , which does not pass through F , intersects l at some point P , which is the image of the

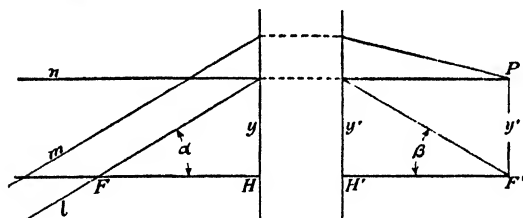


Fig. 57.

source of these rays. From the diagram, it is evident that $y = f \tan \alpha$, where $f = FH$. Similarly the ray n , parallel to the axis in object space, cuts the axis at F' at an angle β in image space.

Therefore $y' = f' \tan \beta$, where $f' = F'H'$. These expressions for y and y' are useful in the experimental location of the principal planes.

440. Combinations of lenses. A pair of thin lenses close together may be treated approximately as a single thin lens. We have seen that such a lens modifies the original curvature according to the relation $C_2 = -C_1 + \phi_1$, where ϕ_1 is the power of the lens. A second lens would act on C_2 (assuming no change of curvature between the lenses) according to the relation $C_3 = C_2 + \phi_2$, where ϕ_2 is the power of the second lens, and C_3 is the curvature of the emergent wave front. Therefore $C_3 = -C_1 + \phi_1 + \phi_2$. But $\phi_1 = 1/f_1$, and $\phi_2 = 1/f_2$; therefore, setting $C_1 = 1/p$, and $C_3 = 1/q$, we have

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f_1} + \frac{1}{f_2} \quad (1)$$

where both f_1 and f_2 may be either positive or negative according to whether the lens considered is converging or diverging.

If the object is at infinity, $1/p = 0$, and the image distance q is the focal length f' of the combination. Introducing these values in (1), we obtain

$$\frac{1}{f'} = \frac{1}{f_1} + \frac{1}{f_2}, \text{ or } f' = \frac{f_1 f_2}{f_1 + f_2}. \quad (2)$$

If we make use of diopters, the calculation is still simpler, for then the power of the combination is given by $\phi' = \phi_1 + \phi_2 = 100/f'$ diopters. But this expression, like (2), is valid only when the lenses are close together.

If the two thin lenses are not in close contact, we must allow for a change in curvature of the wave front as it advances from the first to the second lens. Let us assume that the first lens produces a converging wave. Then the radius of curvature of the wave front C_2 , as specified in Article 431, is decreased in passing from the first to the second lens. In Fig. 58, let r_2 be the radius of the curvature C_2 , and let d be the distance the wave front travels between the lenses. Then r_3 is equal to $r_2 - d$. But $r_2 = q_1$, the image distance from the first lens, and this image is a *virtual* object for the second lens; therefore, if we apply the lens equation to the second lens, we must substitute $p_2 = -(q_1 - d)$, and obtain

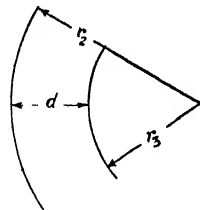


Fig. 58.

$$\frac{1}{-(q_1 - d)} + \frac{1}{q_2} = \frac{1}{f_2}. \quad (3)$$

But q_1 is found from the same equation applied to the first lens, or

$$\frac{1}{p_1} + \frac{1}{q_1} = \frac{1}{f_1}. \quad (4)$$

Therefore, to locate the final image, solve equation (4) for q_1 and substitute in (3).

If, for instance, $f_1 = f_2 = 20$ cm, $d = 10$ cm, and $p_1 = 30$ cm, then from equation (4) $q_1 = 60$ cm, and $q_1 - d = 50$ cm, and from equation (3)

$$\frac{1}{q_2} - \frac{1}{50} = \frac{1}{20},$$

and $q_2 = 14\frac{2}{3}$ cm, measured from the second lens.

If the lenses had been close together, by using equation (1) we should obtain $q = 15$ cm.

SUPPLEMENTARY READING

- R. A. Houstoun, *Intermediate Light* (Chap. 4), Longmans, Green, 1925.
——, *A Treatise on Light* (Chap. 3), Longmans, Green, 1930.
J. Valasek, *Elements of Optics* (Chap. 7), McGraw-Hill, 1928.
J. P. C. Southall, *Mirrors, Prisms and Lenses* (Chap. 7), Macmillan, 1933.

PROBLEMS

1. A micrometer microscope, sighted on an object at the bottom of a jar, must be raised 0.77 cm when a layer of 2 cm of carbon bisulphide covers it, in order to restore the focus. What is the index of refraction of the liquid? *Ans.* 1.63.
2. Calculate the power of a double convex lens in diopters when its index of refraction is 1.5, and the radii of curvature of its faces are 18 cm and 24 cm. *Ans.* 4.85 diopters.
3. Calculate the focal length of a meniscus lens having the same radii as the lens of Problem 2. *Ans.* 144 cm.
4. A lens forms at 60 cm a real image of an object at 45 cm. What are its focal length and its power in diopters? *Ans.* 25.7 cm; 3.9 diopters.
5. A converging lens whose focal length is 40 cm forms a real image at a distance of 60 cm. How far is the object from the lens? *Ans.* 120 cm.
6. If the object is at a distance of 16 cm from a converging lens whose focal length is 24 cm, locate the image. *Ans.* Virtual; 48 cm from the lens.
7. A diverging lens has a power of 4 diopters. How far from the lens is the virtual image of an object at 50 cm? *Ans.* 16.7 cm.
8. Locate the optical center of a meniscus lens whose maximum thickness is 1.5 cm and whose radii are 12 cm and 20 cm. *Ans.* 0.56 cm.
9. A candle flame is focused on a screen 180 cm away from it, by a converging lens. The lens is then moved 45 cm nearer the screen and forms a new and smaller image there. What is the focal length of the lens? *Ans.* 42.2 cm.
10. In a pair of thin lenses close together, one is diverging with a focal length of 84 cm; the other is converging with a focal length of 60 cm. The combination is 3 m from an object. What is the distance between object and image? *Ans.* 10 m.
11. If the diverging lens in Problem 10 is moved 40 cm away from the converging lens, which remains 3 m from the object, what is the distance between image and object? *Ans.* 4 m.
12. Two converging lenses, whose focal lengths are 40 cm and 24 cm, are 80 cm apart. An object is 60 cm from the 40 cm lens. How far is the image from the 24 cm lens? *Ans.* 15 cm.
13. If the object in Problem 12 is 30 cm from the 40 cm lens, how far is the image from the 24 cm lens? *Ans.* 27.3 cm.
14. A compound lens is made of a converging lens of 12 cm focal length and a diverging lens of 15 cm focal length placed 3 cm behind it. What is the back focal length of the combination? *Ans.* 22.5 cm.

CHAPTER 34

Optical Instruments

441. The photographic camera. This device has its name from the earlier "camera obscura" (dark room), a light-tight enclosure fitted with a lens which formed a real inverted image of external objects on a wall opposite the lens. In photography the image is formed on a plate sensitive to light. The portions of the plate illuminated during the exposure undergo chemical changes which are later made visible in the process of "developing," after which the "fixing" bath removes all sensitivity.

In general, the object lies far beyond the doubled focal length of the lens, so that the image is formed between f and $2f$. It is therefore inverted and smaller than the object. If the object is far distant, the plate is in the plane of the principal focus. But for nearer objects, the distance between lens and plate must be lengthened, and if the object is at $2f$, the image distance is the same, and the photograph is the same size as the object. This arrangement is used in copying cameras to reproduce drawings or manuscripts in the same scale as the originals. Enlarging cameras and those used in microphotography have the object just outside of the principal focus, and the image is enlarged as much as is desired.

442. The projection lantern. This is an ordinary camera reversed. That is, object and image have changed places. Instead of forming a small image, close to the lens, of a large object much farther off, the lantern forms a large distant image of the illuminated slide close to the lens. In both the camera and the projection lantern, the ray construction and lens equation are identical, being simply the case of a real image formed by a converging lens, as illustrated in Fig. 48.

In Fig. 59 is shown the arrangement of the various parts of a projection lantern. Light rays from an arc A are brought together by a condensing lens C . This gives an even illumination over the slide P placed just outside of the principal focus of the projecting lens. This lens is compound, with principal planes H and H' as

indicated, and it forms a real image of the slide at a distance q on the screen S .

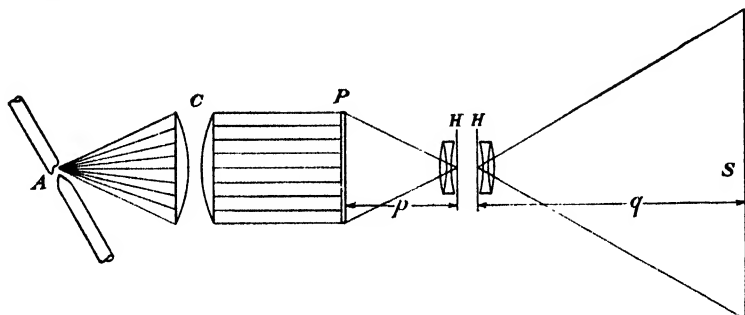


Fig. 59.

443. Rating of camera lenses. A rating of the “speed” of a lens depends upon its opening, or the size of stop used. It is measured in terms of the ratio of the focal length to the diameter of the stop or lens opening. This is proved as follows: Illumination varies as the area of the luminous source (other things being equal) and inversely as the square of the distance between the source and illuminated surface. That is, $I \propto A/S^2$, where A is the area of the stop, or effective lens opening, and S is the distance from that opening to the plate. But A varies as the square of the diameter of the opening, and S for distant objects may be taken as the back focal length of the lens combination. Therefore

$$I = \frac{kd^2}{f^2},$$

where k is a constant.

The time of exposure varies inversely as the illumination, so that the speed of a lens depends upon f^2/d^2 , and the ratio f/d may be taken as its measure. Or rather, f/d is the measure of the speed with an opening whose diameter d may be that of the unstopped lens or of some diaphragm. Thus if f is 12 inches and d is $\frac{3}{4}$ inch, then the lens is said to have an aperture of $f/16$, since the focal length is sixteen times the stop diameter. In order to halve the exposure, we should need a stop having twice the area, or a diameter of $\frac{3}{4}\sqrt{2}$ inches. The aperture f/d is then 11.3, written $f/11.3$. Similarly an aperture of $f/8$ would call for one quarter of the exposure time needed by $f/16$. Thus if we know the correct exposure for a plate of known sensitivity when the aperture is, say, $f/16$, the exposure for other openings is easily computed quite independently of the particular make of lens employed.

444. The telephoto lens. Photographs of distant objects may be enlarged from the original negative. But such enlargements usually are somewhat lacking in detail and clearness, so it is often desirable to have a larger image on the original negative. This may be done by using a very long focus lens and then greatly extending the back focus of the camera by an extension bellows. But a better way is to use a telephoto lens combination which enlarges the image without changing the back focus.

A telephoto combination is made by placing a diverging lens of considerable power behind a converging lens of less power and nearer the converging lens than its principal focus. Thus a real image is formed, as was explained in Article 426. If the focal lengths of the two lenses and their separation are correctly chosen, the principal planes are located well outside of the positive lens, as shown in Fig. 60.

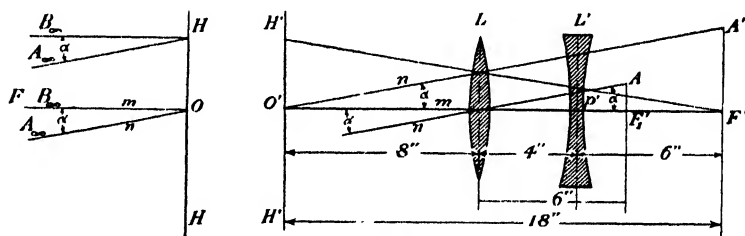


Fig. 60.

This is constructed† to scale for a double convex lens having a 6-inch focus placed 4 inches in front of a double concave lens having a 3-inch focus. The line H is the trace of the first principal plane, if we assume the light traveling from left to right. The line H' is the trace of the second principal plane, and O and O' are the principal points. The point F' is the back principal focus of the combination. Calculation shows that the equivalent back focal length $O'F'$ is 18 inches, and that the image $A'F'$ formed there is 6 inches behind the negative lens L' . This is the same distance it would be behind the positive lens L used alone, as indicated by the image AF_1' . Therefore the relative sizes of $A'F'$ and AF_1' is 18:6, or 3:1. We have thus obtained a magnification of three diameters without altering the back focus $P'F'$ of the camera.

445. The eye. This is really a camera obscura, with the screen, or **retina**, concave toward the lens, so that objects away from the axis

† J. P. C. Southall, *Mirrors, Prisms and Lenses* (pp. 368 ff.), Macmillan, 1933.

are brought to a sharper focus than is possible on a plane surface. The lens L , shown in Fig. 61, is known as the **crystalline lens**. It is doubly convex with its outer surface less curved than the inner one. This lens lies between A , the **aqueous humor** (really another lens), and V , the **vitreous humor**, both of which are more or less fluid and have different refractive indices. The **iris**, II , is a diaphragm having a circular opening of variable diameter, so as to control the amount of

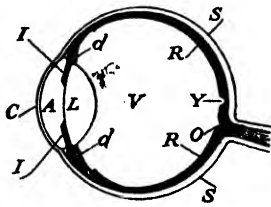


Fig. 61.

light admitted to the retina, R , where the image is formed. The outer coatings of the eye are C , the transparent and tough **cornea**, which protects the lens, and S , the opaque, horny **sclerotic**. Between the sclerotic and the retina is a thin black membrane called the **choroid**, which, like the black interior of a camera, prevents internal reflections.

Unlike the photographic camera, focusing is accomplished mainly by altering the focal length of the lens. This is done by the muscles dd , which act upon the crystalline lens so as to increase chiefly its front curvature, enabling it to focus on nearby objects. When the eye is relaxed, the lens is in a condition for focusing distant objects on the retina. But when nearer objects are viewed, the focal length is decreased by thickening the lens as a result of muscular effort, so that the image distance remains nearly constant for a wide range of object distances. This process, known as **accommodation**, is more perfect with young persons, who can see objects distinctly even five or six inches away, than with older people. In general, 10 inches, or 25 centimeters, is regarded as the *distance of most distinct vision*. This distance, denoted by D , is taken arbitrarily as the minimum object distance for the normal eye.

The image received on the retina is inverted, but the sensation produced in the brain, where it is carried by the optic nerve, O , is interpreted as an erect visual image.

On the axis of the lens is a yellowish area of the retina (Y), which is somewhat thicker there than elsewhere. This is known as the **yellow spot**. At its center is a depression into which only very fine nerves extend, and where our ability to see the minute structure of objects is greatest. The yellow spot in general, and its center (**fovea centralis**) in particular, is that portion of the retina where the particular thing we are looking at is focused, and visual impressions there are more distinct than elsewhere. Objects seen "out of the corner of the

eye" are focused at other points on the retina, but are not brought to our attention with the same intensity.

The place where the optic nerve enters the eye is not adapted to receiving visual impressions directly, and is known as the **blind spot**. We are not aware of this gap in our field of vision because it is filled in with the color of its surroundings. However, it may be found by placing two black dots side by side and 7 or 8 centimeters apart on a piece of white paper. Then if the left eye is closed and the left-hand dot is viewed with the right eye at the distance of most distinct vision, the other dot disappears.

446. Defects of the eye. In some eyes the lens has too short a focal length compared to its distance from the retina, which may be unusually great owing to a very deep eyeball. Such an eye cannot focus clearly on distant objects, although a sharp image may be formed of objects much too near to be seen clearly by a normal eye. This is nearsightedness or **myopia**. The image of an object at a distance is formed at F , which lies in front of the retina, as shown in Fig. 62(a), but with the aid of a diverging lens it may be made to fall directly on the retina, as in Fig. 62(b).

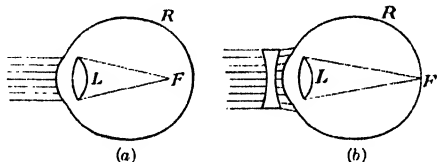


Fig. 62.

(b). If such an eye cannot focus on an object more distant than 6 centimeters, for instance, the focal length of the proper concave lens to enable it to see clearly at any specified distance, as 24 centimeters, may be found as follows: For the unaided eye,

$$\frac{1}{6} + \frac{1}{q} = \frac{1}{f}. \quad (1)$$

For the eye and eyeglass combined,

$$\frac{1}{24} + \frac{1}{q} = \frac{1}{f} + \frac{1}{x}, \quad (2)$$

where x is the focal length of the required lens. Then subtracting (2) from (1), we obtain

$$\frac{1}{6} - \frac{1}{24} = -\frac{1}{x}, \text{ whence } x = -8 \text{ centimeters.} \quad (3)$$

This indicates a negative lens having a power in diopters of $-100/8 = -12.5$, and the eye may be regarded as having an excess curvature of $+12.5$ diopters.

When the lens of the eye has too long a focal length, due either to too little curvature or to an abnormally shallow eyeball, the image of a nearby object is formed behind the retina, as in Fig. 63 (a), though more distant objects are seen clearly. Such an eye must be aided by a converging lens, which adds, as it were, the necessary curvature to

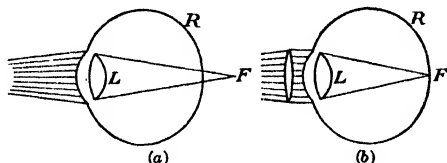


Fig. 63.

the lens of the eye, as shown in Fig. 63 (b). This defect is called **hypermetropia**, or farsightedness. The requisite lens for reading at the standard distance of 10 inches may be found by the

same method as that just used for myopic eyes. Suppose a farsighted person cannot see objects distinctly when they are nearer than 3 feet. Then for the unaided eye

$$\frac{1}{36} + \frac{1}{q} = \frac{1}{f}. \quad (1)$$

But with the added convex eyeglass,

$$\frac{1}{10} + \frac{1}{q} = \frac{1}{f} + \frac{1}{x}. \quad (2)$$

Subtracting (2) from (1) as before, we obtain

$$\frac{1}{10} - \frac{1}{36} = \frac{1}{x}, \quad (3)$$

whence $x = 13.8$ inches, or 35.1 centimeters. This indicates a positive lens having a power of 2.85 diopters, which is needed to correct a deficiency of -2.85 diopters.

Either problem could also have been solved by considering it to be the function of the eyeglass to form a virtual image at the distance required by the eye. In the case of the farsighted eye, this would mean a lens which would form a virtual image at 36 inches of an object at 10 inches. Then, since virtual image distances are negative, we have

$$\frac{1}{10} - \frac{1}{36} = \frac{1}{x}, \quad (4)$$

which is the same as equation (3) above.

Loss of the power of accommodation is a third defect of the eye. This difficulty, known as **presbyopia**, is common to most persons after the age of about 45, and is due to decreasing power of the muscles to thicken the crystalline lens so as to enable it to focus on

nearby objects. It differs from hypermetropia in that the lens and depth of the eyeball may be perfectly normal, but for reading or any other close work, converging lenses must be used to supply the curvature which the muscles are unable to furnish.

447. Astigmatism of the eye. Many persons suffer from a defect of the crystalline lens in which cylindrical curvature is added to its normal spherical curvature. This is known as **astigmatism**, a name derived from the Greek, which refers to the fact that the lens does *not* form a *spot* ($a = \text{not}$, *stigma* = spot) image of a point object.

The thickening of the glass tube of a clinical thermometer above the thread of mercury is a well-known example of a cylindrical lens. The width of the thread normal to the axis of the cylinder is magnified, while its length, parallel to the axis, is not. Thus a cylindrical lens acts like a spherical lens in forming the image of a line normal to its axis (that is, the *width* of the thread of mercury), but it has no lens effect upon a line parallel to its axis. Lines having intermediate directions are of course more or less affected according to their inclination.

We may then imagine an astigmatic eye to be made up of a positive spherical lens L , combined with a positive cylindrical lens L' , whose axis, let us assume, is horizontal, as shown in Fig. 64 (a). Light from

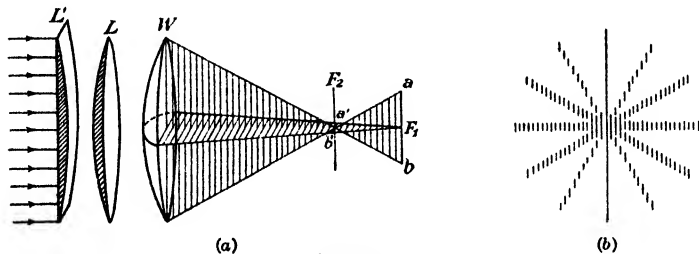


Fig. 64.

a distant point source emerges from the lens combination having a converging wave front W which converges more rapidly in a vertical than in a horizontal plane, because L' acts as a lens in the vertical plane and not horizontally. Consequently the point source has a short vertical line image ab in the focal plane F_1 , and a short horizontal line image $a'b'$ in the focal plane F_2 . If the retina lies in the F_1 plane, the only case we shall consider, vertical lines are seen distinctly, while lines differently directed are blurred, horizontal lines the most so, as shown in Fig. 64 (b). If L' is rotated about the axis of the eye, the distinct line assumes a corresponding direction, though always normal to the axis of the cylinder.

To correct ocular astigmatism, we may place a plano-concave cylinder L'' (not shown) in front of the combination L'/L which represents the eye in Fig. 64 (a). If L'' has the same curvature as L' but of opposite sign, L'' neutralizes the effect of L' , and F_2 coincides with F_1 , forming a point image on the retina. As astigmatism is usually associated with myopia or hypermetropia, cylindrical curvature is usually combined with spherical in the glasses prescribed. Suppose, for instance, that the myopic eye of Article 446 has a vertical cylindrical curvature which increases its excess power from 12.5 diopters in the horizontal plane to 15 diopters in the vertical plane. There are several possible prescriptions for such an eye in which spherical and cylindrical curvatures are combined in one lens. One of them is to use two crossed cylinders of powers -15 diopters vertical, and -12.5 diopters horizontal. Or we might use a negative spherical lens of power -15 diopters combined with a positive horizontal cylinder of power $+2.5$ diopters. A third solution is to combine a spherical lens of power -12.5 diopters with a negative vertical cylinder of power -2.5 diopters.

448. The simple microscope. Any converging lens may be used as a simple microscope to make objects appear larger when seen as virtual images. But to be of much value the lens must have a short focal length. Sometimes two or three such lenses close together take the place of one, though the principles involved are the same, and the focal length of the combination, to be regarded as a unit, is found in the manner explained in Article 440.

The magnification of such a lens or lens combination is the ratio of the apparent size of the object seen through it, to that of the same object

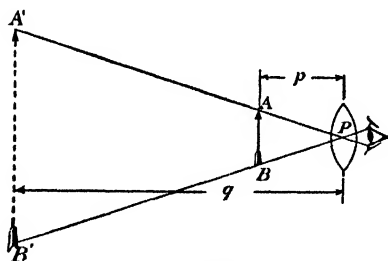


Fig. 65.

seen at the distance D of most distinct vision. A comparison of the observed sizes of the object really means a comparison of the angles it subtends at the eye in the two cases; hence if the eye were at the optical center of the lens in Fig. 65, both the object and virtual image would appear of the same size. This would be almost true

with the eye very close to the lens, in the position indicated, so that from one point of view there is no magnification. However, the distance p , which is equal to or less than f , is less than 10 inches when

the lens has a small focal length as supposed. Therefore the same object seen by the naked eye at 10 inches would necessarily subtend a smaller angle than APB . Thus we may regard a simple microscope as a device enabling the eye to see objects distinctly at much shorter distances than would otherwise be possible, and so making them appear larger. If the angles we are considering are small, we may measure them approximately by the ratio of the length of the object to its distance from the eye, in the two cases to be compared. Therefore the angle subtended when the object is seen by the naked eye at the distance D is approximately l/D radians, where l is the length AB ; but when seen through the lens with the eye as close as possible, the angle is approximately l/p radians. Now p may be made equal to f , which means that the image is at infinity and that the rays entering the eye from each object point are parallel to each other. This puts the least possible strain on the eye, though the magnification is slightly reduced, and its value is then equal to $l/f \div l/D = D/f$. But if the object is a little nearer the lens, so that the virtual image is formed at the distance of most distinct vision, the magnification, $D:p$, when we apply the usual lens formula $1/p = 1/D + 1/f$, is $D(1/D + 1/f) = 1 + D/f$, a slightly larger value than the preceding one.

449. The reading glass. If the eye is not close to the lens, we are using it like a "magnifying" or "reading glass," and the calculation is more complicated. Let S in Fig. 66 be the distance between the

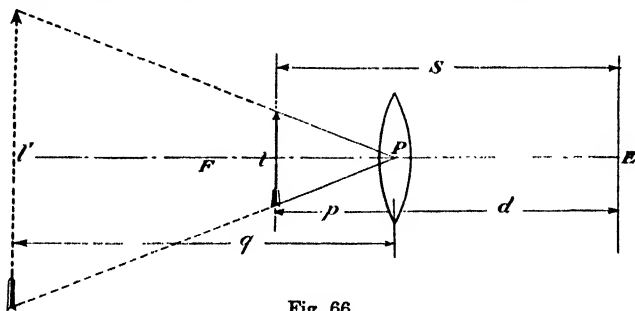


Fig. 66.

object and the eye, and d the distance between the lens and the eye. Then the magnification based on a comparison of the apparent size of the object at the same distance, with and without the lens, is given by $M = l'/(q + d) \div l/(p + d)$. But $l'/l = q/p$, and from the lens formula, $q = pf/(f - p)$. Substituting and reducing, we have

$$M = \frac{f(p + d)}{f(p + d) - pd} = \frac{fS}{fS - pd}. \quad (1)$$

If S is considered as constant, which is generally approximately the case in using a reading glass, the magnification depends upon f and the product pd . Evidently M is a maximum for a given value of f (that is, a particular lens) when pd is a maximum. This occurs when $p = d$, because when the sum of two variables is constant, as in this case ($S = p + d = \text{constant}$), their product is a maximum when they are equal. But p cannot exceed f in forming a virtual image; therefore the above condition can be realized only when $f \geq S/2$. This is usually the case, as most reading glasses have a focal length a little greater than half the usual distance from the reader's eye to his book. Then the glass should be held midway between the book and the eye to obtain the maximum magnification at that distance.

450. The compound microscope. For magnifications larger than are possible with a single lens or lens set, we must resort to two lenses or two lens combinations, each of which contributes to the enlargement. The object whose length may be called l is just outside the principal focus of the smaller lens L , known as the **objective**, shown in Fig. 67. This lens (really a combination of several lenses) has a

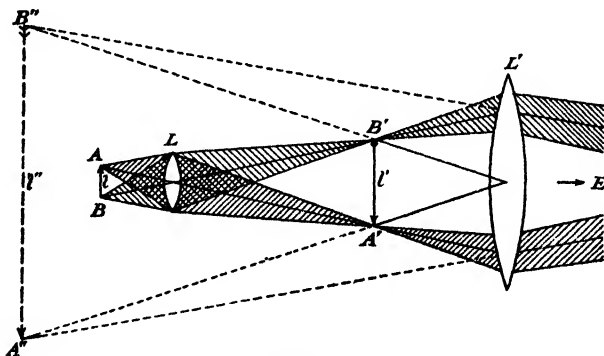


Fig. 67.

very short focal length f , and the real image l' is much larger than the object, though not very far away from it. This real image is formed at or just inside of the principal focus of the other lens combination L' , known as the **ocular**, of focal length f' , and the virtual image, whose length is l'' , is still larger than l' .

In determining the magnification, we shall refer to the simplified diagram, Fig. 68, where only the undeviated rays passing through the optical centers of the lenses are needed. The total magnification M is the product of the magnifications due to the two lenses, or M_1M_2 .

But $M_1 = q/p$, and $M_2 = 1 + D/f'$ if the ocular is focused for maximum magnification. Therefore

$$M = \frac{q}{p} \left(1 + \frac{D}{f'} \right), \quad (1)$$

where D is the distance of most distinct vision at which the virtual image $A''B''$ may be formed.

Since the object is very near F_1 , p equals f approximately. Also, since $A'B'$ is near F_2 , as shown, $q = \Delta + f$ approximately, where Δ , known as the "optical tube length," is the distance between the focal points F'_1 and F_2 . Therefore, substituting these values in (1),

$$M = \frac{(\Delta + f)(f' + D)}{ff'}. \quad (2)$$

This is the maximum magnification. But if the real image is formed at F_2 to avoid eyestrain, $M_2 = D/f'$ and $q = \Delta + f$, exactly. Then, setting $p = f$ as before, and substituting these values in (1), we obtain $M = (\Delta + f)D/ff'$, or, dropping f from $\Delta + f$ as small compared to Δ ,

$$M = D\Delta/ff'. \quad (3)$$

In equation (3) there are three independent variables which determine M . Of these only the optical tube length can be altered at will. It is usually about 6 inches, but may be as long as 10 inches.

451. The reflecting astronomical telescope. The two chief types of this instrument are known as the Gregorian and Newtonian telescopes. The former is rarely seen, but the latter is still used.

In the Newtonian type, a concave mirror M of silvered glass forms a real image of the distant object, as shown in Fig. 69. This would lie across the axis of the telescope but for the totally reflecting prism

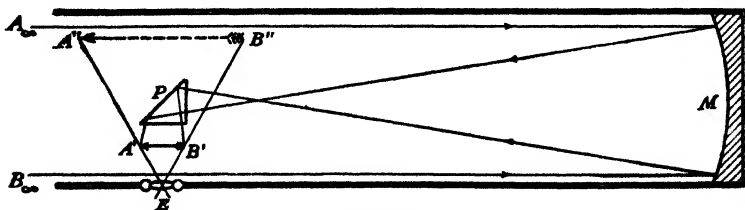


Fig. 69.

P , which causes the image to be formed at $A'B'$ near the eyepiece built into one side of the telescope near its open end. This lens forms a magnified image at $A''B''$ whose size depends upon the focal length of M and of the ocular.

The magnification is the quotient of these focal lengths, because the image $A'B'$ of a distant object would be formed at the principal focus

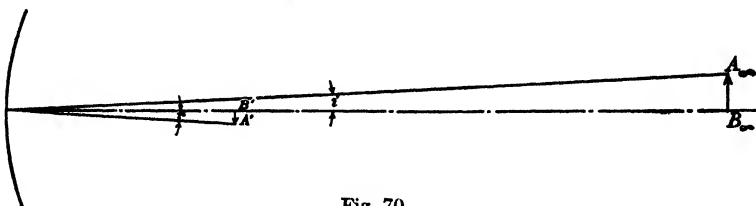


Fig. 70.

of the mirror if there were no prism to reflect it. Therefore it subtends at the mirror an angle of $A'B'/f$ radians equal to the angle subtended by the object itself, as is readily seen from Fig. 70, where the object (AB) subtends the angle i , and the image $B'A'$ subtends an equal angle r . The angle subtended by the virtual image $A''B''$ (Fig. 69) at the eye is approximately $A'B'/f'$ radians, where f' is the focal length of the ocular, so that the ratio of the angles is $f:f'$, as stated above. In the great hundred-inch parabolic reflector at Mt. Wilson, the prism is replaced by a mirror which reflects the light back toward M . Then near M the light is reflected sideways, as in Fig. 69. Thus the ocular can be located conveniently near the base of the instrument, and the telescope is only about half as long for the same degree of magnification.

452. The refracting astronomical telescope. Most modern telescopes are of this type, which was invented by Kepler. The objective

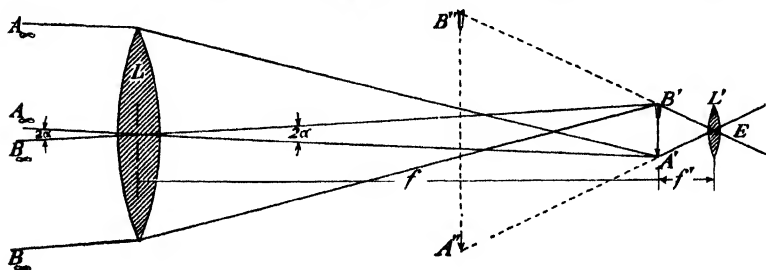


Fig. 71.

L in Fig. 71 forms an inverted real image $B'A'$ at a distance f , so that the angle subtended at the lens by both image and object is $B'A'/f$

radians. The ocular forms the still inverted virtual image $B''A''$, and if $B'A'$ is at the principal focus of L' , the angle subtended at the eye is approximately $B'A'/f'$ radians; therefore the magnification is $f:f'$, and the length of the telescope is $f + f'$. It is then obvious that high magnifying power demands a long tube to allow for a long focal length of the objective.

453. The terrestrial telescope. When a telescope is used to examine the heavenly bodies, the inversion by the astronomical instrument

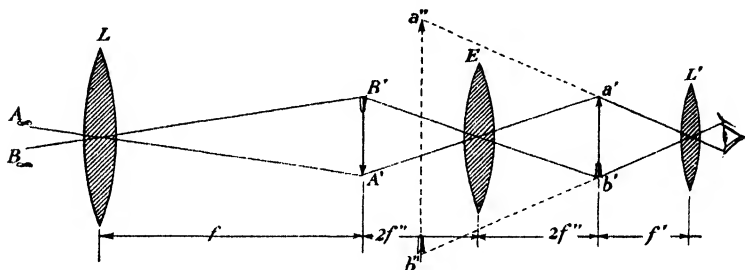


Fig. 72.

is of no consequence, but when objects are viewed in a landscape, it is necessary to have an erect image. This could be accomplished in the simplest way by using a third, or *erecting*, lens E between objective and ocular, as shown in Fig. 72. If E is placed twice its focal length from the real inverted image $B'A'$, it forms a reinverted image of the same size at $a'b'$. The distance between $B'A'$ and $a'b'$ is equal to $4f''$, where f'' is the focal length of E . The virtual image $a''b''$ formed by the ocular L' is then erect and magnified according to the ratio $f:f'$, as in an astronomical telescope. But since this instrument is longer by $4f''$, its total length is $f + f' + 4f''$.

In actual practice, E consists of two similar lenses distant $2f''$ from each other, with the image at the principal focus of one of them, as shown in Fig. 73. In this way an inverted image of the same size as the object is obtained, and the total increase of length is $4f''$, as in the case of a single lens. But there is a gain in illumination, because $B'A'$

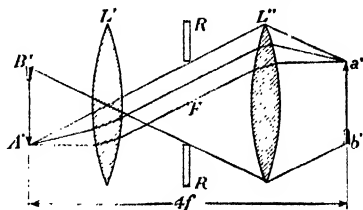


Fig. 73.

is twice as near L' as it would be from the single erecting lens E of equal focal length, while there is no loss of light between the lenses, as the rays are parallel there. A diaphragm R is placed midway be-

tween the lenses, where the rays from the extreme limits of the field cross each other, one purpose being the elimination of stray light reflected from the walls of the telescope.

454. Galileo's telescope. The earliest form of telescope was invented by Galileo in 1609. It enabled him almost immediately to discover the satellites of Jupiter. Their periods of revolution were observed soon after by Kepler who used, at first, the same instrument. This gave added confirmation to the Copernican theory, and to Kepler's first and second laws of planetary motion, published the same year.

Today Galileo's telescope survives almost exclusively in the opera glass having a magnifying power of from 2.5 to 3 diameters.† For this purpose it is simpler and more compact than any other glass, but if higher magnification is needed, Galileo's telescope has serious drawbacks, which have led to its being abandoned for astronomical use. The chief of these is the fact that when designed for high magnification, its field is much more restricted than that of an ordinary telescope.

The instrument consists of a fairly large converging objective L , in Fig. 74, and a smaller diverging ocular L' . A distant object

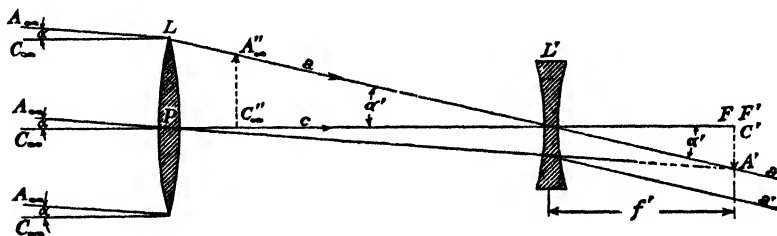


Fig. 74.

$(AC)_\infty$ would form a real image $A'C'$ if the lens L' were removed. This image may be regarded as a virtual object for L' , as explained in Article 426. If it is at the principal focus F' of L' , then p equals $-f'$, and the lens equation reads

$$-\frac{1}{f'} + \frac{1}{q} = -\frac{1}{f''}$$

whence $q = \infty$. That is, there is formed in object space a virtual and erect image $(A''C'')_\infty$ whose extreme rays enter the eye as a

† This means that the *linear* dimensions of the object appear so many times larger. Its *area* however is multiplied by the square of M .

parallel beam aa' . The magnification is α'/α , and as the two principal foci are assumed to coincide, evidently $\alpha = A'C'/f$, and $\alpha' = A'C'/f'$ radians. Hence $M = f/f'$. The length of the telescope is $f - f'$, and if the magnification is three diameters, as constructed in Fig. 74, the distance between lenses is only twice the focal length of the eyepiece.

455. The prism binocular. This ingenious instrument combines the compactness of Galileo's telescope with the higher magnifying power and field of the terrestrial glass. The magnification is accomplished by the same arrangement of lenses as that used in the astronomical telescope, but the formation of an erect image is brought about by two pairs of total reflections within two prisms, each reversing the direction of the beam by two ninety-degree deviations.

A single reflection from a plane mirror results, as we have seen, in either perversion or inversion of the image, so that it takes two such processes both to pervert and invert. Now the real image formed by a converging lens is both perverted and inverted, but in the prism binocular, two pairs of reflections restore it to its natural condition for examination under the eyepiece. At the same time, the actual length of the instrument demanded by a given magnifying power is greatly re-

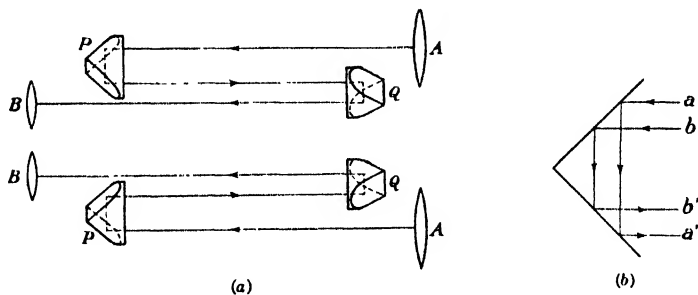


Fig. 75.

duced. This is shown in Fig. 75 (a), where the lenses AA are the objectives, and BB are the oculars. The effective distance between them (approximately the focal length of A) is not far from three times the length of the instrument.

The two internal reflections in the prisms P result in neutralizing the *perversion* of the real image formed by A , while the two internal reflections in the prisms Q reinvert the image so that it is seen erect. The reversal by the two internal reflections of each prism is made clear from an inspection of Fig. 75 (b), where the incident rays are

reflected in reverse order. If a is above b , b' is above a' . But if a is to the right of b as seen from the source, then a' is to the left of b' as seen from the same point. Therefore prisms P pervert, and prisms Q invert.

SUPPLEMENTARY READING

Hardy and Perrin, *The Principles of Optics* (Chap. 11), McGraw-Hill, 1932.
J. P. C. Southall, *Mirrors, Prisms and Lenses* (Chap. 13), Macmillan, 1933.
S. H. Sage, *The Microscope*, The Comstock Publishing Co., 1932.
L. Bell, *The Telescope*, McGraw-Hill, 1932.

PROBLEMS

1. It is desired to form a real image of an object magnified 12 diameters, with a lens whose focal length is 36 cm. What are the object and image distances? *Ans.* $p = 39$ cm; $q = 468$ cm.

2. A camera lens has a focal length of 10 in. and a diameter of 2.5 in. When wide open, the correct exposure with a certain plate and illumination is $1/800$ sec. What is its maximum rating? What exposure would be proper with a stop rating $f/22.6$? *Ans.* $f/4$; $1/25$ sec.

3. The distance between a slide in a projecting lantern and the screen is 4 m. The lens has a focal length of 36 cm. How far from the slide should it be placed? *Ans.* 40 cm (or 360 cm, impracticable).

4. A converging lens forms a virtual image of an object magnified 6 diameters. The object is 18 cm from the lens. What is the focal length of the lens? *Ans.* 21.6 cm.

5. A myopic eye cannot see objects clearly beyond 10 in. What is the focal length of the lens needed to enable it to see objects clearly at a distance of 6 ft.? At infinity? *Ans.* 11.61 in.; 10 in.

6. A farsighted eye cannot see clearly objects nearer than 4 ft. What is the focal length of the lens needed to enable it to see clearly objects at a distance of 14 in.? *Ans.* 19.8 in.

7. Calculate the magnifying power of a simple microscope focused to form a virtual image at 10 in., if the focal length of the lens is 1.5 in. *Ans.* 7.7 diameters.

8. What is the magnifying power of a reading glass whose focal length is 10 in., if it is held 6 in. from a book with the eye 1 ft. away from the lens? *Ans.* 1.7 diameters.

9. What is the maximum magnification attainable with the lens of Problem 8, and with the eye at the same distance from the book? *Ans.* 1.8 diameters.

10. A compound microscope has an objective whose focal length is 5 mm, and an ocular whose focal length is 2 cm. What is the maximum magnification when the lenses are 10 cm apart? *Ans.* 216 diameters.

11. What is the magnification of the microscope of Problem 10, focused to avoid eyestrain, when f is not assumed short compared to Δ ? What is the approximate value when f is assumed negligible? *Ans.* 200 diameters; 188 diameters.

12. A compound microscope has an objective of 0.5 cm focal length and an ocular of 0.8 cm focal length. How far apart should they be to obtain a magnification of 500 diameters, avoiding eyestrain? *Ans.* 8.8 cm.

13. An opera glass 8 cm long has an objective whose focal length is 10 cm. What is its magnifying power? *Ans.* 5 diameters.

14. The length of a refracting astronomical telescope is 5 m. Its magnifying power is 249 diameters. What is the focal length of the ocular? *Ans.* 2 cm.

15. An astronomical telescope has an objective of 120 cm focal length and an eyepiece of 2 cm focal length. The lenses are 152 cm apart for observing an object with minimum eyestrain. How far off is it? *Ans.* 6 m.

16. An astronomical telescope is adjusted to form a real image of the sun 15 cm beyond the ocular. The objective has a focal length of 150 cm and the ocular has a focal length of 2 cm. How far apart are the lenses? *Ans.* 152.3 cm.

17. How far apart should the lenses be in Problem 16 when the telescope is used to observe the moon with minimum eyestrain? *Ans.* 152 cm.

18. What is the focal length of the objective of an astronomical telescope with a distance of 84 cm between the lenses and a magnifying power of 20 diameters? *Ans.* 80 cm.

19. The objective of a terrestrial telescope has a focal length of 80 cm. The erecting lens has a focal length of 18 cm, and the total length of the telescope when observing distant objects is 156 cm. What is its magnifying power? *Ans.* 20 diameters.

20. An opera glass measures 3 in. between objective and ocular. The focal length of the objective is 5 in. What is its magnifying power? *Ans.* 2.5 diameters.

CHAPTER 35

Dispersion and Spectra

456. Dispersion of white light. The bending of a beam of light when it passes through the interface between two different media has been called refraction. But this word is derived from the Latin prefix *re*, + *frangere* (to break), and was probably taken as the name of the phenomenon already described, because white light when refracted is *broken* down into the various colors of which it is composed. As some colors are bent more than others, this decomposing process is now called **dispersion**. It is due to the fact that light of different wave lengths travels with different speeds through a refracting medium. This means that the medium has a different refractive index for the different wave lengths (colors), instead of a single value, as we have so far assumed.

If a narrow cylindrical pencil *P* of white light meets an optically denser medium, as in Fig. 76, it is no longer a cylinder after passing

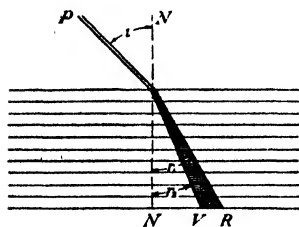


Fig. 76.

the interface, but broadens out into a wedge of light, and when it falls on the lower face of the slab shown in the diagram, the spot formed there is elongated and colored, beginning with red farthest from the normal *N*, and ending with violet nearest *N*. This dispersion is so slight as to be barely noticeable with a slab of glass or shallow layer of water, but may be made increasingly evident by

increasing the depth, or by using solids or liquids of especially high dispersive power.

It is evident from the figure that the refracting medium has a higher refractive index for violet light than for red, since the violet is bent more than red toward the normal. This means that violet light is more retarded than red in its passage through such a medium. In general, light of shorter wave length is more retarded in passing through transparent bodies than light of longer wave length, and has a higher refractive index.

457. Wave length of light. As in all other forms of wave motion, the velocity of light is equal to the wave length times the frequency, ν , or $V = \nu\lambda$.† When light is retarded by a medium such as glass, the frequency is unaltered, but since V decreases, λ must decrease also in direct proportion. Therefore λ is not a constant quantity for a given vibration, but depends upon the medium. In general, values of λ for different colors are given as the wave length in air or in a vacuum. They are so short for visible light that several special length units shorter than the centimeter are commonly employed to measure them. These are the **micron**, μ , which is defined as a millionth of a meter, or 10^{-6} cm; the **millimicron**, $m\mu$, which is a thousandth part of a micron, or 10^{-7} cm; and the “tenth meter,” or **Ångström unit**,‡ defined as 10^{-10} m, or 10^{-8} cm. Thus the waves emitted by common salt volatilized in the flame of a Bunsen burner have an average length of 5893 Å, 589.3 $m\mu$, or 0.5893 μ , according to the unit chosen. As the velocity of light in air is nearly 3×10^{10} cm/sec., the frequency of these particular waves is $3 \times 10^{10} / (5893 \times 10^{-8}) = 5.09 \times 10^{14}$ vibrations per second.

458. The chromatic spectrum. The dispersion of white light into a series of colors may be considerably increased by using two refracting surfaces instead of the one considered above. When the second

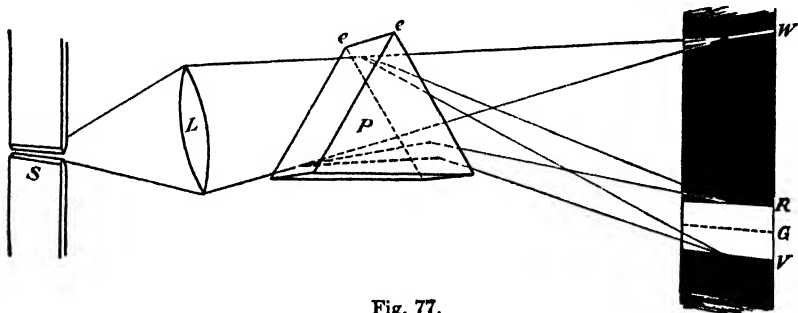


Fig. 77.

surface forms a prism out of the medium, it increases the deviation, as proved in Article 424, and the colors are further separated at the same time. This arrangement is shown in Fig. 77. A source of white light, not shown, illuminates the narrow horizontal slit S , and the lens L , before the prism is introduced, forms a real white image

† In conformity with current practice, we shall use the Greek letter ν (nu) to represent the frequency of light waves, instead of n or f .

‡ This unit is named after the Swedish physicist, Anders J. Ångström (1814–1874). It is pronounced *awngstrum* in Swedish.

of the slit on the screen at W . But when the prism is placed in the path of this convergent beam, instead of a single image of the slit, there is a succession of colored images, forming a continuous band from red to violet, shading gradually into each other. This series of images of the slit forms what is known as the **chromatic** (colored) **spectrum** or simply the **spectrum**, of the white light of the source.

To obtain a good spectrum, the refracting edge ee of the prism must be accurately parallel to the slit, and the prism should be rotated about this as an axis until the resulting spectrum shows minimum deviation. In a darkened room the spectrum is then clearly seen with the colors arranged in the order of red, orange, yellow, green, blue, and violet. There is no sharp limit to the various colors, each shading into the next by imperceptible degrees. Thus yellow is followed by greenish yellow, yellowish green, and green. Between blue and violet is a shade sometimes described as indigo, which is not quite blue, and not quite violet. Purple, as that name is commonly used, and magenta, do not appear at all, for they are not pure spectral colors, but mixtures of the two end colors, red and violet, in different proportions.

If the light which illuminates the slit is monochromatic, that is, light of a single wave length, there is no dispersion, and a real image of the slit is formed where the corresponding color belongs in the continuous spectrum. The line G , in Fig. 77, would be a luminous green image of the slit on a dark background if the slit were illuminated with monochromatic green light. This shows clearly that the band of spectral colors is really a continuous succession of slit images, as has been stated.

If any color or single wave length is missing from the incident light, a black band or line occurs at the corresponding part of the spectrum, and this may be described as a missing group of images, or of a single image of the slit. If the slit is narrow, the line corresponding to a given wave length is narrow also. In short, *spectral lines*, as they are called, are *slit images*, whether they are luminous or dark lines. They are straight if the slit is straight, curved if it is curved, and would not be lines at all if the light came through a pinhole.

459. Recombination of the colors. In a celebrated experiment by Newton, two prisms of the same kind of glass and having the same refracting angle were arranged as in Fig. 78. In this way a pencil of white light is broken up into colors by the first prism. Then, in passing through the second, the colors are recombined so that the emergent pencil is white again, having the same direction as the inci-

dent pencil. It is as if the light had passed through a single glass slab included between the parallel planes ab and $a'b'$, as indicated by the dotted lines.

This experiment proved that white light is really made up of all the colors of the spectrum into which the first prism resolved it, and that these colors are not created by the process of refraction. Therefore, true white light contains all possible wave lengths between the extreme limits of the red and violet ends of the spectrum, as perceived by the human eye. Wave lengths which correspond to a greater bending than that of violet, or less than that of red, are, strictly speaking, not light at all, although they are emitted by most luminous sources. At any rate, they are invisible and contribute nothing to the whiteness of light.

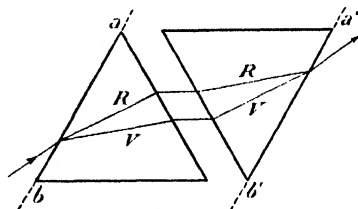


Fig. 78.

460. The spectrometer. This is an instrument used for the production, examination, and measurement of spectral lines. It is shown in plan in Fig. 79, and consists of a *collimator*, C , at one end of which is a slit S illuminated from the source whose rays are concentrated there by the lens L , which is not a part of the spectrometer.

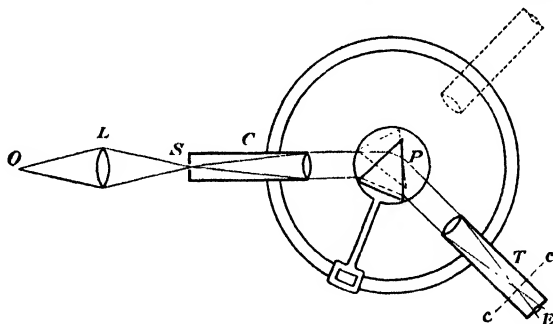


Fig. 79.

The slit is at the principal focus of a lens at the other end of the collimator tube, so that the emergent rays are parallel. These enter the prism P which is mounted on a table that can be rotated about a vertical axis (that is, normal to the plane of the diagram). The objective of the telescope T receives the rays which are brought to a focus in the principal plane of the ocular. In this plane are located the cross hairs cc , so that the spectrum image and cross hairs lie in the

same plane and are simultaneously magnified by the ocular. The collimator is fixed, but the telescope and prism rotate about the same axis, while verniers attached to them read angles on a circular scale commonly graduated in degrees and half degrees.

461. Continuous spectra. Incandescent solids and liquids emit light of all wave lengths up to a minimum length which depends upon the temperature. Thus red-hot iron emits a spectrum that is continuous as far as it goes, but it does not reach appreciably into the blue or violet, while white-hot iron would show all the colors. A jet of illuminating gas emits a continuous spectrum caused by incandescent solid particles in the flame. Incandescent liquids also emit a continuous spectrum, and if gases under very high pressure could be heated to incandescence, they might also. At ordinary pressures, gases cannot be made luminous by temperatures available in the laboratory, but they are made to give out light in other ways, such as by an electrical discharge, and then the spectrum produced is not continuous.

462. Line spectra. If an electric discharge passes through a partially exhausted tube containing the gas whose spectrum is desired, only a selected group of wave lengths is radiated, and the spectrum appears as a series of bright-colored line images of the slit. These lines serve as a valuable means of identifying gases, and this method

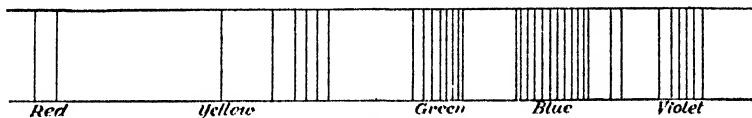


Fig. 80.

of identification is known as **spectrum analysis**. Such a series of lines is shown in Fig. 80, which represents the **line**, or **emission spectrum**, of oxygen.

Substances not normally gaseous may be made to emit line spectra under certain conditions, as when their salts are volatilized in the flame of a Bunsen burner, or when an electric arc passes between the terminals of a metal whose emission spectrum is desired. In the case of an ordinary spark discharge, the resulting spectrum has lines due both to the metal of the electrodes and to the surrounding gas. But as this gas is progressively exhausted, the lines due to the electrodes disappear, and there remain only those due to the gas as described above. In general, both arc and spark spectra have many more lines than those produced in a flame.

463. Band spectra. These spectra have a very complicated structure of lines grouped in "bands," each band being densely packed with lines at its edge of longer wave length, and fading out with wider distances between the lines in the direction of shorter wave lengths, as indicated in Fig. 81. The bands are of varying breadth and spacing, depending upon the substance producing them.

Band spectra are emitted under certain conditions by many substances which normally give line spectra. If a spark discharge is

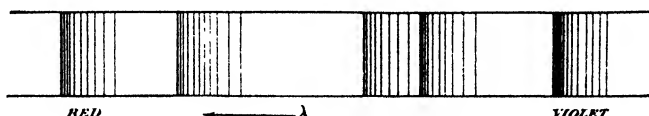


Fig. 81.

passed between electrodes coated with a compound of mercury, for instance, a feeble discharge produces a band spectrum, while during a heavy discharge, only the line spectrum is emitted. Since a heavy discharge disrupts the molecules of a compound, this is interpreted to mean that band spectra are characteristic of the molecule, while line spectra belong to the atom.

Some compounds break down so easily, as a result of a discharge, or in a flame, that it is very difficult to obtain their band spectra. This is why, in the Bunsen flame, the salts of the alkaline elements give only the line spectrum of their metallic component, so that sodium, whether as a chloride, carbonate, or nitrate, gives the same flame spectrum, the characteristic yellow lines. Modern theories of atomic and molecular structure have been guided to a large extent by attempts to give a rational explanation of the nature of the spectra emitted by various substances under various conditions. A detailed study of band spectra has made it possible to learn a great deal about the way in which atoms are bound together in molecules.

464. Absorption spectra. When white light passes through certain liquids or solids containing dissolved metals, the spectrum is discontinuous, having broad black bands not very sharply defined at their edges. These bands indicate what is known as selective absorption. Ruby glass, such as is used in the photographic dark room, absorbs all wave lengths except for red and a little orange. Blue cobalt glass has several bands due to the absorption of all the orange and yellow, and most of the red and green, leaving blue and violet, and narrow bands of red and green. Didymium glass transmits all colors freely but yellow, where a strong and well-defined absorption band is found.

A solution of chrome alum has strong absorption bands in the violet, orange, and yellow. This makes it appear green by transmitted daylight, but purple by the light from an incandescent lamp, which is relatively rich in red.

Some substances transmit light very freely up to a certain limiting wave length, and are then opaque. Glass ceases to be transparent for waves shorter than 3500 \AA , which is only a little beyond visible violet (about 4000 \AA), and it is also quite opaque in the extreme infrared region of thermal radiation. On the other hand, quartz is highly transparent in the ultraviolet region as far as 1800 \AA , while rock salt is particularly transparent in the infrared. Other substances are quite opaque in the visible spectrum and highly transparent to thermal radiations, as was pointed out in Article 297.

Selective absorption occurs also as a result of reflection from the surfaces of crystals. It is by no means complete after a single reflection, but if a beam of light undergoes repeated reflection from successive surfaces of the same kind of crystal, only very long wave lengths are left. These are usually called by the German name of *rest strahlen* (residual rays). Rubens and Nichols in this way isolated waves as long as 0.01 mm .

465. The Fraunhofer lines. It was explained in Article 300 that good radiators are good absorbers of radiant energy. This is true in a very special sense in the case of selective absorption, for then light of precisely those wave lengths that a body absorbs are readily emitted when that body is heated to incandescence. Substances which emit a line spectrum under suitable conditions also absorb the particular frequencies emitted, and no others, exactly as a given piano string responds to a musical tone by absorbing some of its energy and then re-emitting it, as has already been pointed out.

In the year 1802, Wollaston noticed some dark lines crossing the otherwise continuous spectrum of the sun. These were rediscovered in 1814 by Joseph von Fraunhofer, who made a careful study of them and mapped about 600, giving the more prominent ones alphabetical names, beginning with *A* in the extreme red, and ending with *H* in the violet. The significance of these lines was not at first understood, but in 1859 Kirchhoff gave the correct explanation based on laboratory experiments in which similar lines were reproduced. The visible surface of the sun, called the **photosphere**, is much hotter than the surrounding gaseous envelopes, known as the **reversing layer** and the **chromosphere**, which constitute the sun's atmosphere. The photosphere behaves like an incandescent solid in emitting a continuous

spectrum. Its light in passing through the reversing layer and chromosphere partially loses by absorption those wave lengths which the less luminous gases are emitting. The energy thus absorbed is in turn re-emitted; otherwise the gaseous envelopes would become steadily hotter. But the re-emission takes place in all directions, so that only a small fraction of the luminous energy that started toward the observer can reach him, and the light due to these particular wave lengths is very much less intense than that which has not been so absorbed. These absorbed portions of the spectrum then appear dark against their more luminous surroundings. However, during an eclipse of the sun, it may be seen that the reversing layer and chromosphere emit bright line spectra, showing that the gases they are made up of are emitting at the same time they are absorbing.

This phenomenon may be reproduced in the laboratory by passing light from a glowing solid through the flame of a Bunsen burner in which some common salt is being volatilized. If the continuous spectrum is weak and the sodium flame is strong, the two yellow lines (D_1 and D_2) appear bright against a darker background. But if the continuous spectrum is made gradually stronger, the D lines gradually fade in comparison, and finally appear reversed, being dark against a much brighter background. This is because the flame absorbs these wave lengths from the continuous spectrum, although they are still emitted by the fainter light of the flame as much as before.

The Fraunhofer lines extend both into the ultraviolet and into the infrared of the solar spectrum, and 22,000 have been mapped. The infrared portion of the solar spectrum has been photographed as far as 11,900 Å, and explored to 20,000 Å by means of the bolometer. The ultraviolet reaches only to about 3000 Å, because shorter waves are absorbed by the earth's atmosphere. Many, though not all, of these lines have been identified with the spectra of known elements. These elements must therefore exist in gaseous form in the solar envelope, or in the earth's atmosphere, whose absorption accounts for some of the lines, such as the B line in the red due to oxygen. The D lines are due to sodium, and another, the D_3 line, seen only as an emission line from the chromosphere during an eclipse, is due to the gas helium in the outer solar atmosphere. Helium was discovered in the sun before it was known on the earth, and named from the Greek word *helios*, meaning *sun*. Among the other important Fraunhofer lines are the E and G lines due to iron, the C and F lines due to hydrogen, and the H and K calcium lines.

The spectra of most of the stars are absorption spectra similar to that of the sun. Many of their lines indicate the presence of known elements, while some are still unrecognized. Some of the diffuse and planetary nebulae have bright line spectra, but the spiral nebulae



Absorption spectrum. Didymium glass.

(a)



Line spectrum. Argon.

(b)



C D E b F G H₁H₂

Absorption line spectrum. The sun.

(c)



Band spectrum. Carbon dioxide.

(d)



Band spectrum. Ammonia.

(e)

Plate 5.

Photographs of spectra arranged to correspond in wave length, with red at the left. In (a) note the remarkably sharp absorption band in the yellow. In (c), the D, E, F, and G Fraunhofer lines are clearly visible. In (d) and (e), the bands look like fluted columns, whence the term "fluted spectrum," by which they were formerly known.

have both types, absorption and emission in the same spectrum suggesting a galaxy of stars like our sun, combined with true nebulous masses to account for the bright lines.

466. Irrationality of dispersion. It would seem natural to expect that the only difference in the distribution of the spectral lines pro-

duced by different prisms would be in their extent, and that the spectrum produced by a highly refrangible (large refractive index) glass prism could be exactly duplicated by a prism of less refrangible glass but with a larger refracting angle. In other words, we should expect, *a priori*, that all spectra of the same substance would be alike except as to extent, and that by proportionally reducing the length of one, or stretching out another, they could be brought into exact coincidence. In Newton's time this was supposed to be the case, but since it was later found untrue, this peculiarity of the spectra formed by glass prisms is known as the **irrationality of dispersion**.

As an example of irrationality, we might compare the solar spectrum as produced by two prisms, one of flint and one of crown glass, having refracting angles such that the total lengths of the two spectra are the same at the same distance. As crown glass has less refracting

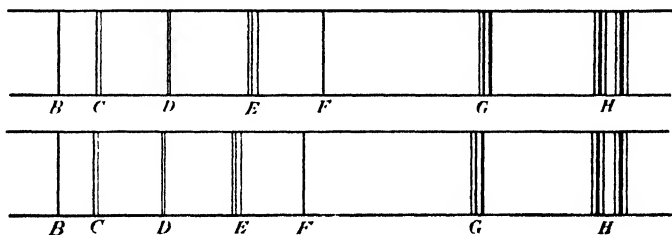


Fig. 82.

power than flint, its prism must have a larger angle to produce an equally long spectrum. But even when this is done, the spectra do not coincide throughout, as is evident from Fig. 82.

On the other hand, we might cut two prisms so that a chosen line, as *F*, might be deviated through the same angle in both spectra. If this is done, the two spectra have different lengths and no other part would be equally bent by both prisms, while if the two prisms have the same refracting angle, their spectra differ in deviation, length, and relative position of the lines.

467. Relative deviation. In order to compare the dispersive properties of different sorts of glass, we must remember that both bending and dispersion depend upon the angle of the prism as well as upon the nature of the glass. A comparison of their ability to bend light is based upon the angular deviation (Fig. 83) of a given line, as D_D , divided by the refracting angle of the prism. This ratio, D/A , may be denoted by δ , and is known as **relative deviation**. Two narrow prisms made of the same glass, but with different refracting

angles, would have the same *relative* deviation for the same line. This follows from equation (2), Article 424, where $n = (A + D)/A$. Solving it for D , we have

$$D = A(n - 1). \quad (1)$$

Hence D varies as A , if n has the same value in different narrow-angled prisms.

The relative deviation may now be obtained from (1), giving

$$\delta = \frac{D}{A} = n - 1, \quad (2)$$

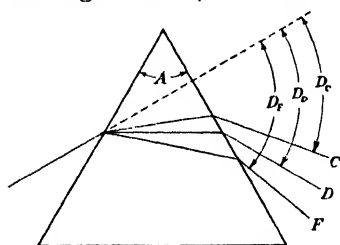


Fig. 83.

where n is the refractive index corresponding to the light that is deviated through the angle D .

468. Coefficient of dispersion and dispersive power. If we wish to compare different sorts of glass with respect to their ability to extend the spectrum, we must choose two lines whose angular separation can be definitely determined. These are usually the C line in the red, and the F line in the blue, both due to hydrogen. The ratio of their angular separation to the refracting angle of the prism is called **relative dispersion**. Thus, with reference to Fig. 83, the angular dispersion exhibited by the F and C lines is $D_F - D_C$, and their relative dispersion is $(D_F - D_C)/A$. This is also called the **coefficient of dispersion**, and may be denoted by Δ .

If we substitute for D_F and D_C their values obtained from equation (1) in the last article, then

$$\Delta = \frac{D_F - D_C}{A} = (n_F - 1) - (n_C - 1),$$

$$\text{or} \quad \Delta = n_F - n_C. \quad (1)$$

This coefficient varies with the nature of the medium, but as explained in Article 466, it does not vary in the same way as relative deviation. A change in one is not accompanied by a corresponding change in the other unless prisms made of the same medium are compared. The ratio between these two properties of a prism (dispersion and bending) is known as **dispersive power**, and is usually denoted by ω . That is, $\omega = \Delta/\delta$. But $\Delta = n_F - n_C$, and if the bending power of a prism is measured by the relative deviation of the sodium (D) lines in the brightest part of the spectrum, $\delta_D = n_D - 1$. Therefore

$$\omega = \frac{n_F - n_C}{n_D - 1}. \quad (2)$$

Sometimes the B and H , or even A and H , lines are used in determining Δ , thus giving slightly different values for the dispersive power.

Equations (1) and (2) are derived on the supposition that the prism has a small refracting angle, and they cannot be applied to large-angled prisms. But actually they are most used in connection with lenses, and then this condition of a small angle A is sufficiently realized.

The following table gives values for the refractive indices n_C , n_D , and n_F , as well as δ_D , Δ_{FC} , and ω_{FC} , for several transparent media.

	n_C	n_D	n_F	δ_D	Δ_{FC}	ω_{FC}
Water (20° C).....	1.3311	1.3330	1.3371	0.3330	0.0060	0.0180
A light crown glass (15° C)	1.5145	1.5170	1.5230	0.5170	0.0085	0.0164
A dense flint glass (15° C)	1.6444	1.6499	1.6637	0.6499	0.0193	0.0297
Rock salt (18° C).....	1.5407	1.5443	1.5534	0.5443	0.0127	0.0233

469. Chromatic aberration. When a ray of light traverses a lens, it is deviated exactly as if it were passing through a prism whose faces were planes tangent to the lens at the points of entry and emergence. Such a ray then must suffer dispersion as well as deviation. If a plane wave originating in a point source is brought to a focus by a thin lens of small aperture, as in Fig. 84, the different colors have different foci. This is because the glass has different indices of refraction for each wave length, and f , as we have seen, depends upon n according to the relation $\phi = 1/f = (n - 1)(S_1 - S_2)$, as shown in equation (1), Article 432. Since n is greatest for violet light, f is shortest for these rays, and longest for red, as indicated in the diagram.

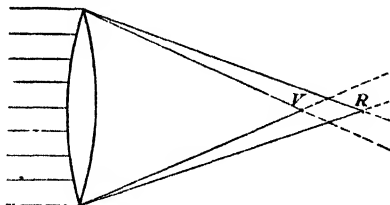


Fig. 84.

If the point source is focused on a screen placed at the focus V , it is not seen as a sharp point image, but as a small disc with a violet-hued center surrounded by a reddish halo. At R , the center is reddish instead, and has a violet-hued halo. Actually the latter is the better position for a sharp focus because the brightest part of the spectrum (yellow) lies much nearer red than violet. But in either case the image is not sharp, and whether it is that of a star or of some point in an object viewed with a field glass or focused on a photo-

graphic plate in a camera, it is seriously lacking in definition. In the case of a star observed through a telescope which does not correct chromatic aberration, the image is distinctly colored. This defect of lenses is wholly different from spherical aberration, and may appear even when the latter has been eliminated by a suitable design of the lens combinations.

470. Achromatic combinations. To correct chromatic aberration, use is made of the irrationality of dispersion. If dispersion were rational, dispersive power would be the same for all kinds of glass, so that a supplementary lens which brought the foci V and R together would at the same time straighten the ray so that no deviation would result. It would then reduce the combination to the equivalence of a parallel-sided slab, as would be the case if two lenses, one concave and the other convex, were combined, as in Fig. 85 (a), when each is

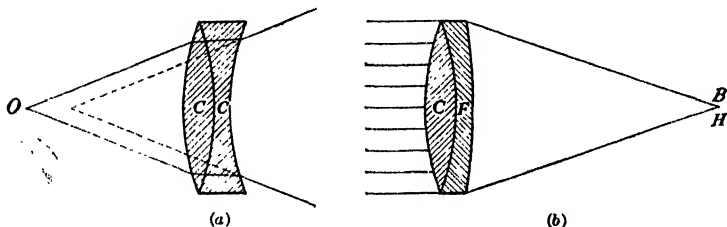


Fig. 85.

of the same kind of glass. But fortunately this is not the case. The problem is to combine two lenses of two kinds of glass so that the two spectra produced may have the same length between two arbitrarily chosen lines. Then the dispersion of one lens will be exactly neutralized by the other at the limits chosen, but there will be an outstanding bending of the light, so that the combination functions as a single lens, as shown in Fig. 85(b) for the B and H lines.

The condition of achromatism is that the focal length of the two lenses for the C and F lines shall be the same. From equation (2), Article 440, $1/f' = 1/f_1 + 1/f_2$, where f' is the focal length of the combination. But by hypothesis, $f'_C = f'_F$; therefore the basic equation for calculating an achromatic doublet may be written

$$\left(\frac{1}{f_1} + \frac{1}{f_2}\right)_C = \left(\frac{1}{f_1} + \frac{1}{f_2}\right)_F. \quad (1)$$

471. The direct-vision spectroscope. It is possible to combine two prisms made of two kinds of glass with different dispersive powers, so that some one wave length will pass through the combination with-

out deviation, while others will be more or less bent, thus forming a spectrum spread out on either side of the undeviated image of the slit. This "dispersion without deviation" is the reverse of an achromatic combination, where we have "deviation without dispersion." The arrangement indicated in Fig. 86 is known as a direct-vision spectroscope, and is both portable and compact. It serves the purpose of rapid qualitative examination of the spectra from different sources, but is not well adapted to quantitative measurements.

The design of the direct-vision prism combination is based on equating the deviations produced by the two prisms having different angles.

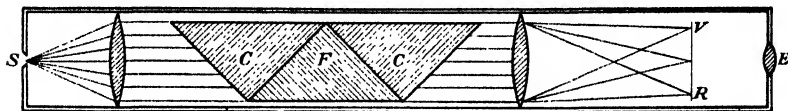


Fig. 86.

The D lines are usually selected to be undeviated, so that in the case of small-angled prisms, $A(n_D - 1) = A'(n_D' - 1)$, from which the ratio of the angles, $A : A'$, may be calculated. In the actual instrument, there are usually two crown-glass prisms and one of flint glass, as shown in Fig. 86, and these are combined with the collimator and telescope in a single tube.

472. Anomalous dispersion. This is a change in the usual order of the colors of the spectrum produced by prisms of certain substances. A portion of the spectrum where the wave length is short has a smaller index of refraction than another portion with longer wave lengths, thus reversing the usual order of the two color groups. This was first noticed by Talbot and later investigated by LeRoux, who, in 1860 and 1861, used a hollow glass prism filled with iodine vapor, and found the violet light less bent than the red. In this case the refractive indices are 1.00205 for red, and 1.00192 for violet.

In 1870, Christiansen, a Danish physicist, investigated the spectrum of a strong alcoholic solution (18 per cent concentration) of fuchsine contained in a hollow glass prism of small refracting angle, and obtained a similar reversal. Fuchsine is one of the aniline dyes, and is often used in making red ink. It is red violet (magenta) by transmitted light, but in the solid state has a bluish-green sheen known as *surface color*, which is due to selective reflection.

The spectra produced by prisms filled with strong solutions of fuchsine, or still better, by thin prisms made of the solid substance, start with blue; then comes violet, then a dark absorption band, and

then red, followed by orange and yellow in their proper order, but at the wrong end of the spectrum. Thus green and part of the blue are missing, and the two transmitted portions are reversed, though keeping to the proper sequence of their component colors, as shown in Fig. 87.

It was later found that a great variety of substances exhibits this phenomenon, including many aniline dyes. Among the number are

<i>R</i>	<i>O</i>	<i>Y</i>	<i>G</i>	<i>B</i>	<i>I</i>	<i>V</i>	<i>Normal</i>
<i>B</i>	<i>I</i>	<i>V</i>			<i>R</i>	<i>O</i>	<i>Abnormal</i>

Fig. 87.

chlorophyll, and glass colored with didymium, uranium, or cobalt. Vapors of metals like sodium, and gases under certain conditions, as

in the protuberances of the sun during an eclipse, also exhibit anomalous dispersion. These all have absorption bands, and it is precisely around such a band that this effect is produced.

Such substances also exhibit selective reflection, though not necessarily as strikingly as fuchsine, and light of any wave length lying within the absorption band undergoes *total internal reflection* at all angles of incidence, including zero degrees. This means an infinite value of the index of refraction, since $n = 1/\sin C$, where C is the critical angle, and at normal incidence, $\sin C = 0$; so $n = \infty$.

473. The Doppler effect. Light, as well as sound, has a Doppler effect. When the source and the eye are approaching or receding from each other, there is an observed change in frequency. A line in the spectrum of a star is displaced toward the red end of the spectrum (corresponding to lower pitch of sound waves) when the star and observer are receding from each other. The same line shifts toward the violet (higher pitch) when the source and the observer are approaching each other. With light, however, there are not two cases, as with sound, and only *relative* motion between eye and source is significant, because we cannot regard space as a fixed "frame of reference" to which motion can be related, such as the atmosphere in transmitting sound waves.

The change in frequency of light, due to the Doppler effect, is synonymous with a change in wave length, because according to the principle of relativity, the velocity of light c , in space, is a constant, and not affected by motion of either the source or the observer. Therefore an increased frequency means shorter wave length, and vice versa. As a consequence of this postulate, it can be proved that

when the relative velocity of *approach* between source and observer is v , the modified frequency ν' is given by

$$\nu' = \nu \sqrt{\frac{c + v}{c - v}}, \quad (1)$$

and for relative recession, it is given by

$$\nu' = \nu \sqrt{\frac{c - v}{c + v}}. \quad (2)$$

These values may also be obtained by taking the square root of the product (geometrical mean) of the observed frequencies for approaching source and approaching observer in a fixed medium, like sound in air, or a receding source and receding observer. Thus the product of equation (2), Article 354, and (3), Article 355, gives $(\nu')^2 = \nu^2 \left(\frac{c + v}{c - v} \right)$ when c is substituted for V , the velocity of sound.

The square root of this expression is equation (1) above.

A measurement of the shift of the spectral lines makes it possible to calculate the speed with which a star is moving in the line of sight, even though it may be thousands of light years away. But its motion at right angles to the line of sight can be found only by the parallax method in the case of a few relatively nearby stars.

474. Spectroscopic binaries. Double stars may also be detected by the Doppler effect, when they are too far off to be resolved into their components by the most powerful telescope. These doublets, known as **spectroscopic binaries**, revolve about an axis lying between the two components. If this axis is not in the line of sight, there will be a component of the motion of each of the two stars in the line of sight. These motions go through a cycle of changes, and are in-

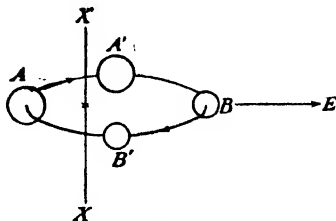


Fig. 88.

dependent of the motion of the binary as a whole, toward or away from the eye. Thus in Fig. 88, the component A of the binary AB is seen just starting toward the eye at E as it rotates about the axis X , while B is about to recede. In these positions there is no Doppler effect, except as the pair may be receding from or approaching the earth. But a quarter of a period later, when they are at A' and B' , the light of a spectral line from A is shifted toward the violet, while a line in the spectrum of B is shifted toward the red. This results in splitting

up a line common to both into two lines that separate to a certain distance, and then meet again at the end of half a period of revolution of the binary. In this way the velocities of rotation and the period may be measured. The period may be as long as several years, but in most cases it is surprisingly short, being less than ten days for the majority, and in a few cases it is only a few hours, a period implying enormous velocities. More than a thousand spectroscopic binaries have been discovered, though in about 80 per cent of them the light from one component is so faint that their double character is inferred from the period of the cyclical displacement of a single spectral line that does not double up in the manner just described.

475. The Zeeman effect. In 1896, Professor Zeeman, of Holland, made the important discovery that when a radiant source emitting a bright line spectrum was placed between the poles of a powerful electromagnet, the lines were split up into a number of component lines. In the simplest case, one line becomes three with the two outer lines equidistant from the original line between them. But if the light is viewed in the direction of the field, a line that yields a triplet in the usual or "normal" Zeeman effect is spread out into two lines symmetrically displaced on either side of its usual position, and to the same distance as the outer lines of the triplet. Lines having originally a more complicated "fine structure" are correspondingly altered, giving rise to very complex patterns.

This linking of radiation with magnetism had already been shown possible by Faraday and Kerr in the field of polarized light, as will be explained later, but Zeeman's discovery gave a clue to the inner structure of the atom which has been of the greatest value. As the effect is found in the spectrum of sun spots, it has also furnished important data on the electromagnetic conditions of solar phenomena.

476. The Stark effect. It was natural to expect that, if a strong magnetic field produces a change in the atomic mechanism concerned with radiation, a strong electrostatic field should produce a similar effect. In 1913, Stark, a German physicist, found this to be the case. The lighter gases, such as hydrogen and helium, are most easily affected, though the phenomenon may be found in the heavier gases also. The result of applying the field is to split up a line into several components, their number depending upon the particular line observed and the strength of the field.

477. Resonance spectra. Professor R. W. Wood discovered in 1906 that if a powerful source of monochromatic light were thrown upon the vapor of metallic sodium, a series of bright lines was emitted

by the vapor, provided the wave length of the illuminating source coincided with one of the lines of that series. This was also found to be the case with mercury and other vapors, and the phenomenon was called **resonance radiation**.

The series of lines so produced is due to absorption by the vapor, which then re-emits the energy absorbed in a variety of frequencies having a well-defined numerical relation to each other. This phenomenon, as well as the Zeeman and Stark effects, is now largely explained by an increasingly intimate knowledge of atomic structure and the theory of quanta, but such explanations are beyond the scope of this volume.

SUPPLEMENTARY READING

R. A. Houstoun, *Intermediate Light* (Chap. 7), Longmans, Green, 1925.

J. K. Robertson, *Introduction to Physical Optics* (Chap. 6), D. Van Nostrand, 1935.

PROBLEMS

1. The strong green line of the mercury arc has a wave length of 0.00005461 cm. Write this number expressed in microns, millimicrons, and Ångstrom units. What is the light's frequency? *Ans.* 5.49×10^{14} v.p.s.

2. Using equation (1) of Article 470, (1) of Article 432, and (2) of Article 468, prove that the condition of achromatism is given by $\omega_1/(f_1)_D = -\omega_2/(f_2)_D$.

3. A positive crown-glass lens has a focal length of 36 cm for sodium light. (a) What is the proper focal length of a flint-glass lens for sodium light to form with the crown-glass lens an achromatic doublet, as specified in Article 470? (b) What is the focal length of the doublet? (Use equation derived in Problem 2, and (2) of Article 440.) *Ans.* (a) -65 cm; (b) +80.3 cm.

4. It is desired to make a doublet of crown and flint glass having a focal length of 48 cm. What are the required focal lengths of the two lenses for sodium light? (Use equation (2) of Article 440 and equation of Problem 2 simultaneously.) *Ans.* +21.5 cm and -38.9 cm.

5. A prism of crown glass has a refracting angle of 10° . What is the required angle of a flint-glass prism to form a direct-vision spectroscope? (Actually such small angles would not be used, but they are here assumed so that the simplified formula of Article 471 may be applicable.) *Ans.* 7.96° .

*6. Calculate the angular width of the spectrum between the *C* and *F* lines formed by the prisms of Problem 5, using equation (1) of Article 467. *Ans.* $4.15'$.

*7. If the emergent light of the prism pair of Problems 5 and 6 is a plane wave, and is then focused on a screen by a lens of 50 cm focal length, what is the length of the spectrum between the *C* and *F* lines? *Ans.* 0.60 mm.

8. By means of a diffraction grating (Article 505), the lines of a certain star's spectrum indicate a 0.2 per cent increase in frequency. How fast are the star and the observer approaching each other? (Take $c = 3 \times 10^{10}$). *Ans.* 599.4 km/sec.

CHAPTER 36

Interference of Light

478. Conditions of interference. We have already seen in Articles 325 and 352 that water waves and sound waves “interfere” when their vibrations are in opposite phase, so that crests are neutralized by troughs, or compressions are neutralized by rarefactions. Something very similar can occur when two beams of light meet in opposite phase, and the result is seen in alternating light and dark regions where we should expect continuous illumination. Such phenomena were first studied extensively by Fresnel, and are convincing evidence of the wave character of light.

Just as in the case of sound or water waves, so with light, we must employ only a single wave length, so that we may obtain simple and stationary interference patterns. Light of a single wave length is called monochromatic. It may be produced in sufficient purity for most purposes by burning fused sodium chloride in the flame of a Bunsen burner, although there are really two sodium lines of slightly different frequency. A more strictly monochromatic source is the mercury arc used with a filter which passes the light of only one of its spectral lines. The brightest of these is one of the green lines ($\lambda = 5460.7\text{\AA}$), and there are filters which transmit practically all of this light, but stop light of the other visible mercury lines. Such an arrangement constitutes the brightest monochromatic source available in the laboratory.

There are two additional requirements for the interference of light waves, which do not apply to waves of sound or water waves. These are that the two beams which are to interfere must start from the same source, and that the difference of path must not exceed a rather indefinite maximum.

479. Nature of light waves. The fact that light from different sources cannot interfere, or rather cannot produce an interference pattern, has an important bearing on the nature of light waves. The evidence both of interference phenomena and experiments with polarized light shows that light waves are transverse vibrations, but very different from water waves, which are also transverse. In water

waves, the vibrating particles move in circles or ellipses whose plane is vertical and in the direction of propagation. Vibrations of light appear to lie in a plane normal to the direction of propagation, and take place in every conceivable direction within that plane.

As light originates in the individual atoms of the luminous body, and as the smallest point source contains millions of atoms, a pencil of light at any point in space or moment of time represents disturbances of all sorts of phases and directions. But, however complicated, each of these may be resolved into components along rectangular axes, and added up, so that there are then only two independent vibrations at right angles to each other. If these vibrations are combined into a single resultant, this resultant may be regarded as representing instantaneously all the complex vibrations from which it was derived. Its phase and amplitude are the resultant phase and amplitude of the beam, and its direction determines a plane that may be called the instantaneous wave plane. Such a plane includes the direction of the beam, and is thus normal to the plane of vibration of a single (imaginary) vibrating particle. The relation of these planes to each other is shown in Fig. 89. Any plane, as AA , is a plane in which the displacements or disturbances occur. It is normal to the direction of propagation V and to the plane of the sine waves SSS , whose instantaneous position in space is determined by the direction of the vector B .

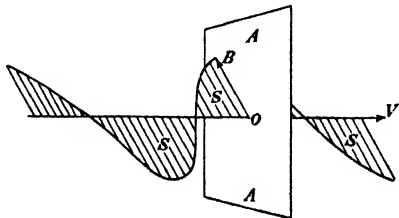


Fig. 89.

From the point of view of the time of a single vibration, the character of vibration of the resultant, represented by B in the diagram, changes very slowly. Thousands of vibrations may elapse before there is any appreciable change in the wave plane, for instance. But from the point of view of *visual impression*, the amplitudes and instantaneous wave planes change very rapidly, as well as the phase relations between two different beams.

From the foregoing it is clear that beams of light from two sources cannot produce an interference pattern. The changes of vibration in one beam, produced by one group of atoms, have no relation to the changes taking place in the other, which originated in another group of atoms. They may indeed interfere from time to time and produce a pattern of bright and dark bands for an instant, but the eye cannot detect such fleeting phenomena. Therefore, in order to

obtain visible interference phenomena, the interfering beams must originate in a single source.

480. Allowable difference of path. The reasoning of the preceding article also makes it clear why two beams originating in the same source cannot interfere if the difference in length of their paths before recombining is too great. It is just possible to detect interference when the path difference is about as great as 78×10^4 wave lengths. Then we can calculate the corresponding distance and maximum allowable time interval for some particular wave length. The *D* lines of sodium have a mean wave length of 5893×10^{-8} cm. Therefore the maximum allowable path difference is $78 \times 10^4 \times 5893 \times 10^{-8} = 46$ cm. The velocity of light is about 3×10^{10} cm/sec.; therefore the time during which the vibrations remain sufficiently constant for interference is $46 \div 3 \times 10^{10} = 1.5 \times 10^{-9}$ second, or about one-and-a-half billionths of a second.

481. Interference from two narrow apertures. In 1801, Thomas Young, an English physician and natural philosopher, discovered the interference pattern produced by light from a "point source" passing through two small holes. This is similar to the case with sound described in Article 352, and is explained as follows:

In Fig. 90, let *S* be a point source of light, such as a small hole in a screen lighted from behind. Let *A* and *B* be two other holes close

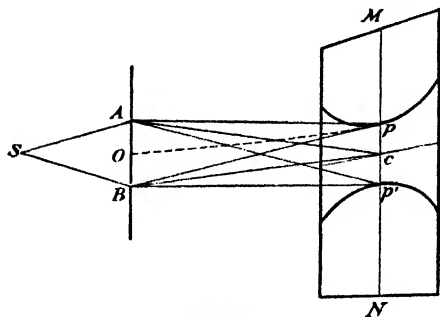


Fig. 90.

together in a screen shown only in section. These two openings, in accordance with Huygens' principle, may be regarded as independent sources of light, and if $SA = SB$, their vibrations are the same in phase, plane, and amplitude at any instant. Then at *C*, or anywhere along the *OC* axis where the distances *AC* and *BC* are

equal, the light arrives from both *A* and *B* in the same phase. The amplitude is then double that due to either source alone, and the energy is quadrupled, because, as in all forms of harmonic vibration, the energy is proportional to the square of the amplitude.

But there are also points like *p* where the difference of path is not zero, but equals half a wave length. Then the vibrations of the two beams at such a point are in opposite phase and destroy each other,

so that the energy developed there is practically zero. The locus of these interference points in the plane of the paper is given by $Bp - Ap = \lambda/2$, which is the equation of the hyperbolic curve shown by the dotted line, as in the similar case with sound waves. This locus is of little interest here, however, but the locus of p (or p' similarly defined) on a screen MN , passing through pCp' and perpendicular to the paper, is of more importance. These loci are hyperbolic also, as shown in Fig. 90, and would appear as two dark curved lines above and below the bright straight-line locus defined by C .

If S , A , and B are slits, seen above in section, whose length (normal to the paper) is greater than the width of the screen MN , the loci

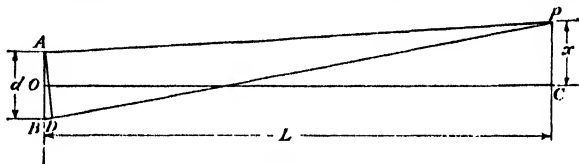


Fig. 91.

of p and p' are straight lines parallel to the bright line of reinforcement through C . But, as we have already learned, at any point p (Fig. 91) where the path difference is an even number of half wave lengths, there is reinforcement, while at a point where the difference of path is an odd number of half wave lengths, there is interference. This results in a succession of bright bands parallel to the slits, and the bands are defined by the relation $Bp - Ap = m\lambda$. They are separated by dark bands whose equation is $Bp - Ap = (2m + 1)\lambda/2$, where m is any integer including zero. The distance x of each band from the line through C is found as follows: Draw the line AD so that $Dp = Ap$. Then if p is in the first dark band above the axis, $BD = \lambda/2$. The triangles pCO and ABD are very nearly similar; therefore $BD/AB = x/L$ approximately, or $\lambda/2d = x/L$. For other dark lines, $(2m + 1)\lambda/2d = x/L$, whence

$$x = \frac{(2m + 1)\lambda L}{2d} \quad (1)$$

This gives the distance of each dark band from the central bright band through C , in terms of the distance between the slits and the distance L between slits and screen. These are easily measured, so that x may be calculated for a given wave length, or the wave length may be calculated if x is measured.

482. Fresnel's biprism. In Young's experiment, the fact that the pattern is rather dim and the bands very close together makes accurate

measurements difficult. But two similar arrangements due to Fresnel give much brighter and broader bands. One of these consists of two mirrors slightly inclined to each other so as to form

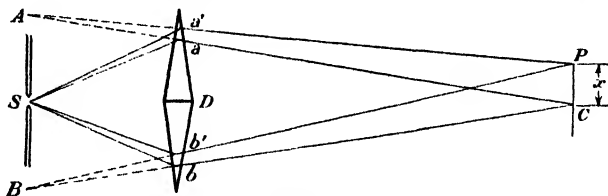


Fig. 92.

two virtual images of the illuminated slit S . Then the reflected beams form an interference pattern such as would be produced by two beams from the virtual images if they were the slits of Young's experiment.

The other arrangement, called Fresnel's biprism, is an arrangement of two narrow-angled prisms, as shown in Fig. 92. The prisms are placed with their bases in contact and parallel to the slit, so that an observer at C on the axis SD produced would see virtual images of S at A and B . The beams of light coming through the two prisms form an interference pattern as if A and B were illuminated slits, as may be understood from the following considerations. Any ray of light from S which reaches C through the upper prism appears to come from A , and has an exactly corresponding ray that comes to C through the lower prism, apparently from B . These symmetrically situated rays reach C after traversing exactly the same distance in air and the same thickness of glass, and so reinforce each other. Therefore all the light which arrives at C from one prism is reinforced by all the light arriving there from the other, and a bright band is formed at the center of the field and parallel to the slit. Points such as P in Fig. 92, lying at a distance x from the axis, receive light from the upper prism

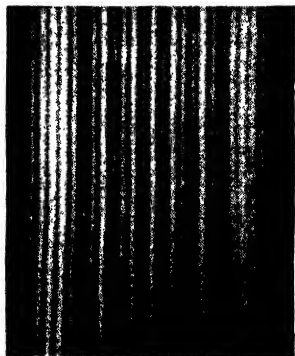


Plate 6.

Photograph of the interference pattern formed by Fresnel's biprism, using a slit source and sodium light. Note curvature.

by a route which is shorter than the corresponding rays from the lower prism, as is evident from the diagram. This results in reinforcement or interference according to whether the virtual path

difference $BP - AP$ (or really $Sb'P - Sa'P$) is an even or odd number of half wave lengths.

483. Colors of thin films. The colors seen in soap bubbles, and in oil films floating on water, are due to interference. The interfering beams are produced by reflection from the two surfaces of the film, and the color is caused by the destructive interference of a portion of the spectrum, leaving the rest to reach the eye. As true white light is composed of all the spectral colors combined in a certain distribution of intensities, any alteration in their distribution, or the removal of any portion, results in coloring the beam. In Fig. 93, let a narrow pencil of parallel rays whose wave front is ab fall upon the thin layer of some refracting medium of thickness t . The ray incident at b is in part refracted at an angle β to d , and then partly reflected at the same angle to c , where it joins that part of the ray through a which under-

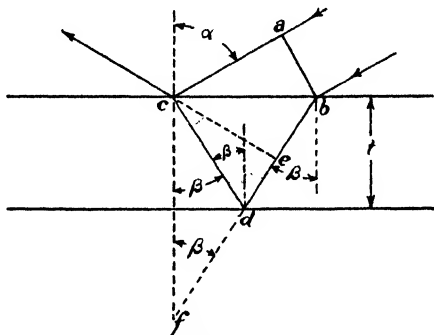


Fig. 93.

goes reflection at c . These are now in a position to interfere, provided their difference of path is an odd number of half wave lengths for some portion of the spectrum.

The geometrical path difference is the difference between the route b to d to c , and that from a to c . Now drop a perpendicular from c to the ray bd , meeting it at e . Since this is perpendicular to the refracted ray, it indicates the refracted wave front of the beam according to Huygens' construction, as shown in Fig. 18, Article 328. Therefore the refracted wave has traveled from b to e , while the wave in air has gone from a to c , and the path difference is edc . But if de is produced to meet at f a perpendicular from c as shown by the dotted lines, then obviously $cd = df$, and the path difference is edf , and $edf = cf \cos \beta = 2t \cos \beta$.

This difference in actual distance is located in the medium of the film, and must be multiplied by the index of refraction of the film to reduce it to an equivalent path in air, so that the path difference is made equal to $2tn \cos \beta$. But it was explained in Article 324 that transverse waves, reflected against dense media back into rare, undergo a reversal of phase; therefore the external reflection at c

results in increasing the effective path difference by half a wave length. The internal reflection at d in the case of oil on water has no reversal, for oil is optically denser than water.

We may now state the conditions of destructive interference for a given wave length as

$$(2m + 1)\frac{\lambda}{2} = 2tn \cos \beta + \frac{\lambda}{2},$$

which reduces to

$$m\lambda = 2tn \cos \beta, \quad (1)$$

where m is any integer.

The preceding theory accounts for the variety of colors seen when oil spreads out over water, because each color depends upon the particular wave length destroyed by interference, and this varies both with the angle of incidence of the light which determines β , and with the thickness of the film.

Soap bubbles and soap films in general exhibit beautiful colors, especially just before they break, when they are thin enough to render m small. The thinnest portion, where the rupture is to occur, finally turns black, because if $2tn \cos \beta$ is less than λ , no interference can take place, even of the first order, except that due to the reversal at the upper surface, which destroys all colors alike. But before this condition is reached, the last color to be seen is a pinkish hue caused by the destruction of violet. This color experiences first-order interference with a minimum thickness, because λ varies as t , and λ is least for violet light.

If soap films are thicker than two or three wave lengths of light, no colors appear, because then the conditions for interference may apply to several wave lengths simultaneously, and the mixture of hues resulting from the removal of several colors from the reflected beam appears as white light.

484. Newton's rings. A classical experiment by Newton illustrates the preceding principle in a very beautiful manner. A plano-

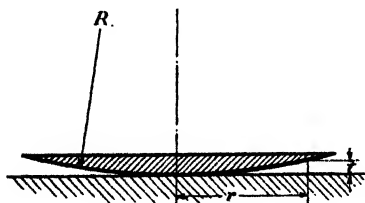


Fig. 94.

convex lens of large curvature is pressed upon a piece of black glass that serves as a mirror, as shown in Fig. 94. In this case the layer of air between the two acts as the medium which produces interference by reflection at the interface between it and the lens, and at the surface of the mirror. Thus the incident light

undergoes two reflections, and the resulting beams can interfere.

The central spot where the lens and mirror touch is black, because of the black-glass mirror and not because of any interference. Then outside this center the first wave lengths to interfere are those of violet light, which leave a pinkish-hued inner ring. Red light is the last to be destroyed in the first-order colors, and the result is a blue-violet outer ring. Then the series begins all over again, but as rings of

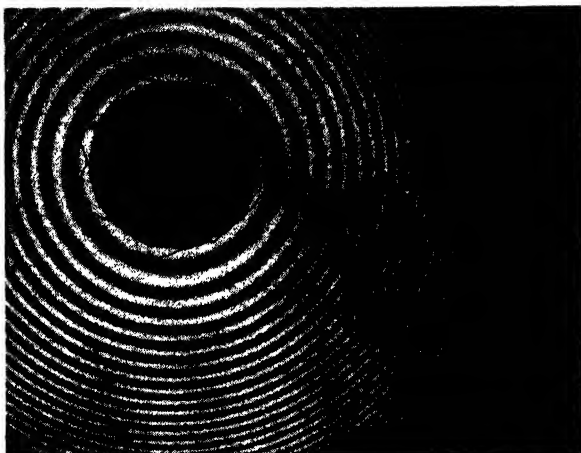


Plate 7.

Photograph of Newton's rings, using sodium light.

higher orders begin to overlap, they are progressively fainter, and soon merge into white light. With monochromatic light a great many black rings are visible, though their position changes with the angle at which they are viewed, and they crowd closer and closer together at increasing distances from their common center.

485. Michelson's interferometer. Perhaps the most remarkable method of producing interference phenomena is that invented by A. A. Michelson, an eminent American physicist. It consists in dividing a beam of light by partial reflection and partial refraction, sending the two beams over different routes to two mirrors, where they are reflected back toward the source and finally recombined to produce interference. This is shown in Fig. 95, where a source of monochromatic light *S* at the principal focus of the lens *L* sends a parallel beam to the so-called "half-silvered" mirror *m*. This has a thin silver coating, indicated by the heavy line, which divides an incident ray *ab* into two of equal intensity. One of these, *a*, the result of reflection at an incident angle of 45° , is bent through 90° and reaches the mirror

M_1 after passing through the transparent slab of glass n . It is then again reflected, and returns to m as indicated by the dotted line a' . There it again divides, but only the refracted portion which reaches the eye at E is to be considered.

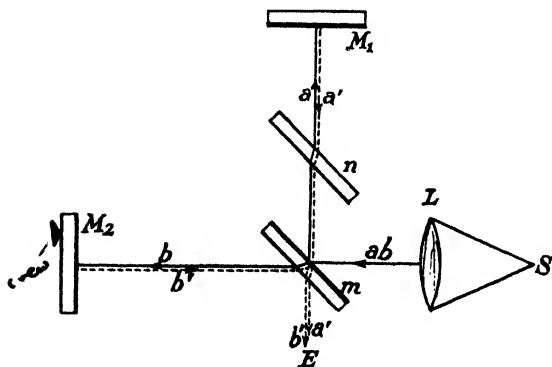


Fig. 95.

The other ray, b , is refracted by the glass slab of the half-silvered mirror, and emerges parallel to its original direction. It is then reflected from M_2 and returns to m , where it passes through the glass and is half reflected internally by the silver film. This second re-

flexion results in a deviation of 90° , and this second ray b' finally reaches the eye parallel with the first, and in a condition to interfere with it.

The slab n is introduced to make the aa' path equivalent to the bb' path, for it will be noticed that the aa' ray passes through the slab m only once, while the bb' ray passes through it three times. Therefore, if n is of the same thickness as m and inclined at the same angle, the distances of the mirrors M_1 and M_2 from the half-silvered surface, which divides and recombines the rays, are equal for equal optical paths of the two rays.

It would seem that if the equality of the ray paths were exact, the two beams should reinforce each other, but this is not the case, for whereas a is externally reflected twice in air against silver, ray b is reflected only once in this way, and once internally in glass against air when it returns to m and is bent downward through 90° , as shown. Therefore ray a has experienced two reversals of phase, and ray b only one, so that there is half a wave length difference of phase between them. This means that they interfere, and the central band of the system thus formed is a dark one.

If one of the mirrors, usually M_2 , is moved so that its surface is always parallel to its original position, but its distance from m is altered, the bands shift across the field and may be counted as they move. If, for instance, M_2 is moved away from m a distance of one millimeter, a path difference of two millimeters is introduced, and

thousands of lines pass before the eye. Since each of them corresponds to a path difference of one wave length of the light used, the total number of bands counted measures in terms of light waves twice the distance through which the mirror is moved. In this way Michelson measured the length of the "meter of the archives" in terms of the wave lengths of three of the lines of the spectrum of cadmium. As measured by the red line in this spectrum, the length of the meter is $1,553,163.6 \lambda$, so that if the standard meter were destroyed, it could be reconstructed from cadmium wave lengths with almost perfect accuracy.

486. The Michelson and Morley experiment. The question of whether or not the velocity of light depends upon the motion of the observer through space began to be discussed quite early in the last century. Certain experiments, such as that by Airy on the aberration of light from the stars, and by Fizeau on the velocity of light through a moving stream of water, gave results which seemed to indicate that the ether, which was supposed to be the vehicle of light, was dragged along with moving matter.

In order to test this question, Michelson and Morley, in 1881, performed a vital and celebrated experiment with the aid of the interferometer described above, though modified to increase the distances of the mirrors M from the half-silvered mirror. It was set up so that one of these distances was parallel to the orbital motion v of the earth through space, and the other at right angles to it.

Now if there is a stationary ether through which the earth is moving, there would be an ether stream, or "drift," which would decrease the velocity of light c , going with it, and increase it when the observer moved against the drift. Then $c - v$ and $c + v$ are the two resultant velocities. One of the rays, as a in Fig. 95, would have the higher speed relative to the apparatus before reflection by M_1 , and the lower speed after reflection (or vice versa), while the other, b , would be affected very much as a rifle bullet would be affected by a wind blowing across its path, causing it to take a longer route in reaching the target.

The difference of time taken by the two rays to complete their separate paths, as influenced by such an ether drift, can be shown to equal lv^2/c^2 , where l is the length of each path. This would mean a shift of the bands from the position they would occupy with no ether drift. Therefore, if the whole apparatus is carefully rotated through 90° , so that the ray b occupies the place of ray a , we should look for a displacement corresponding to twice the time difference, or $2lv^2/c^2$.

Calculation shows that the predicted shift should be easily observable provided l is long enough. But Michelson and Morley found no such effect, although they could have detected one several times smaller than the calculated value.

This negative result was so important that the experiment has been repeated many times under increasingly rigorous conditions, and with longer and longer distances between the mirrors, though always with the same result, except in one instance. Professor Dayton C. Miller, working at the Mount Wilson observatory in California, obtained a small shift of the bands which seemed to indicate a relative motion of the earth, or solar system, with respect to some fixed medium. But the evidence against such a view is so strong that the scientific world decidedly inclines to the belief that it is impossible to detect absolute motion through space, and according to the doctrines of relativity, such an idea as absolute motion is meaningless. Indeed, relativity owes its origin to the Michelson and Morley experiment.

SUPPLEMENTARY READING

Hardy and Perrin, *The Principles of Optics* (Chap. 28), McGraw-Hill, 1932.
T. Preston, *The Theory of Light* (Chap. 7), Fifth Edition, Macmillan, 1928.
R. W. Wood, *Physical Optics* (Chap. 6), Macmillan, 1934.
R. A. Houstoun, *A Treatise on Light* (Chap. 26), Longmans, Green, 1930.
A. Einstein, *Relativity*, Peter Smith, 1931.

PROBLEMS

1. A screen is 1 m from two slits illuminated by sodium light (mean $\lambda = 5893 \text{ \AA}$). The slits are 0.1 mm apart. Calculate the distance of the sixth dark band from the axis. *Ans.* 3.24 cm.
2. What is the distance between the slits in Fig. 91, when the fourth dark band, formed by light of wave length 0.6 microns, is 8 mm from the axis on a screen 2 m away? *Ans.* 0.525 mm.
3. If the dark bands formed by two slits (Fig. 91) are 0.6 mm apart, when $d = 0.4 \text{ mm}$ and $L = 50 \text{ cm}$, calculate λ . *Ans.* 0.48μ .
4. A parallel beam of sodium light strikes a film of olive oil floating on water. The oil's refractive index is 1.46. When viewed at an angle of 30° from the normal, the eighth dark band ($m = 8$) of the system is seen. What is the film's thickness? *Ans.* 1.72μ .

CHAPTER 37

Diffraction

487. Examples of diffraction. It is fairly easy to observe diffraction phenomena with no other apparatus than the two hands. Hold one hand at arm's length toward a window, and obtain a narrow band of light shining between two fingers. Then look at this band through a similar very narrow slit formed by the fingers of the other hand held close to the eye. If the two slits are nearly parallel, the illuminated slit appears wider than when looked at directly, and on either side of it, in the "shadows" of the fingers, are parallel rows of bright and dark *diffraction bands*.

Another experiment is to view an illuminated pinhole H in a screen, through another pinhole H' in another screen two or three feet from the first. If the eye is at certain definite distances from H' , depending on the size of H' , the illuminated hole H appears brighter than it does when seen directly with no screen interposed, while from other positions of the eye it is hardly visible at all.

488. Cause of diffraction. Diffraction phenomena are those in which light is observed to "bend" around corners and produce interference patterns that are not accounted for on the simple assumption that it travels in straight lines. Huygens' principle, as we have so far used it, is insufficient to account for diffraction, and it must be extended as follows: Let da in Fig. 96 be an infinitesimal area of the spherical wave front originating in S . This element sends out a wavelet which we must now regard as effective, not only in the direction of the normal N , as has already been assumed, but also in any direction D , making with N an angle θ that is less than 90° . We shall further suppose that the amplitude of the disturbance in the D direction decreases gradually until it is zero in the direction V normal to N . This supposition is based on observation, and is described as the effect of increasing obliquity. If then such a wave front meets a screen

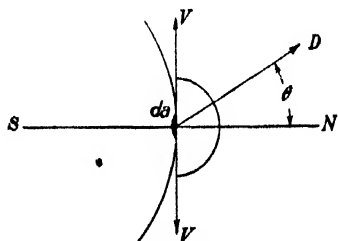


Fig. 96.

having a pinhole whose area is da , the wavelet shown above starts from that hole as a new source, and spreads out in all directions beyond the screen, but with diminishing intensity as θ increases.

489. Half-period elements. In Fig. 97, let MON be the plane section of a spherical wave front originating in the point source S . For an eye at E , the point O , where the axis SE intersects the wave front, is the *pole* of that surface. With O as a center, describe the small circle mn , like a parallel of latitude such that the distance from m to E is half a wave length longer than the distance from O to E ; that is, $mE - OE = \lambda/2$. Similarly the small circle $m'n'$ is defined

by $m'E - OE = \lambda$. The zones mOn and $m'mnn'$ are called **half-period elements**, because each elementary circle of one zone has a corresponding circle in the next, which sends out disturbances differing in phase by half a period at E . Thus each zone interferes almost completely with its neighbor at that point with

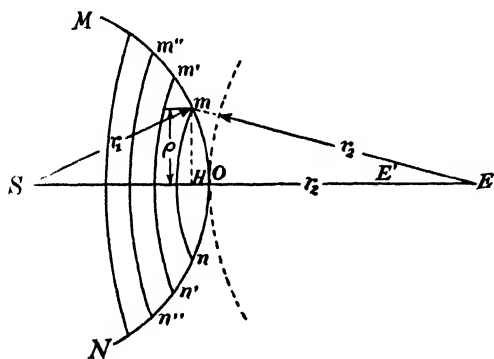


Fig. 97.

reference to which the zones were determined. The zones for an eye at E' would be different, and they vary also with changing values of the wave length.

In most cases the portion of the wave front which we have to consider in problems of diffraction is so small that it may be regarded as substantially plane. It is then easy to prove that the areas of the different zones in a given case are all equal to each other.

490. Resultant illumination. The half-period elements, both of plane and spherical wave fronts originating in a point source, all have practically the same area, provided r_1 and r_2 are large compared to the wave length. We should then expect that they would each produce nearly equal illumination at E , so long as their distances from E do not materially increase. But actually, quite apart from a very gradual increase in distance, there is another and much more important cause of decreasing illumination at E from elements increasingly far from O . This is the effect of increasing obliquity, already referred to. The result is that light from zones increasingly distant

from O becomes rapidly less effective at E , so that only a relatively small number of zones need be considered. If this were not so, light would bend around corners much more than is the case.

We may now express the amplitude of the disturbance at E as the sum of the disturbances due to the series of half-period elements. These are of progressively diminishing amplitude, and as they are alternately in opposite phase, the successive amplitudes have opposite signs. Setting D as the resultant amplitude at E , and d_n as a component due to the n th zone, we have

$$D = d_1 - d_2 + d_3 - d_4 + d_5 - \dots d_n,$$

where $d_1 > d_2 > d_3 > d_4 > \dots d_n$. It can be shown that the limiting sum of such a series is half of the first term. Therefore the amplitude at E , due to the entire wave front, is $d_1/2$, or half of the amplitude that would be produced there by the central zone alone. As wave energy varies as the square of the amplitude, the luminous intensity would be four times as great with only the central zone exposed.

491. Effect of a perforated screen. If a screen having a hole just large enough to transmit the first half-period element were interposed between the wave front and E , we should realize the condition described in the last article, and the point source would appear four times as bright as when seen with nothing intervening. If the source is far enough away to consider the wave front plane, we may calculate the position of E as follows: Referring to Fig. 97, let mH (denoted by ρ) be the radius of the opening in the diaphragm, and let S be so far away that the wave front is a plane defined by mH . Let a denote mE , and let b denote HE . Then $\rho^2 = a^2 - b^2$, and the area A of the opening is given by

$$A = \pi(a^2 - b^2). \quad (1)$$

By hypothesis, $a - b = \lambda/2$, or, $a = b + \lambda/2$.

Substituting in (1), we have

$$\begin{aligned} A &= \pi[(b + \lambda/2)^2 - b^2] \\ &= \pi(b\lambda + \lambda^2/4). \end{aligned}$$

But $\lambda^2/4$ is vanishingly small compared to $b\lambda$, and may be dropped. Then $A = \pi b\lambda$. But A also equals $\pi\rho^2$; therefore $\pi\rho^2 = \pi b\lambda$, and $b = \rho^2/\lambda$, or more generally,

$$b = \rho^2/n\lambda, \quad (2)$$

where n is any integer.

We are now able to calculate b for any assumed pinhole diameter and light wave length. Thus, taking a diameter of one millimeter, $\rho = 0.05$ cm. Take $\lambda = 5 \times 10^{-5}$ cm (a reasonable average), and let n be unity for the boundary of the first order maximum. Then $b = 25 \times 10^{-4} / (5 \times 10^{-5}) = 50$ cm. The various maxima and minima follow as the eye approaches the pinhole. The first minimum is found when $n = 2$: then $b = \rho^2 / 2\lambda = 25$ cm. The second maximum, less intense than the first, is at $b = \rho^2 / 3\lambda = 16\frac{2}{3}$ cm from the opening. The second minimum is at 12.5 cm, and so on to maxima and minima of higher orders.

492. Zone plates. If every alternate zone is eliminated, it is possible to obtain a much greater illumination at E than the maximum due to the central half-period element. Then with, say, only the odd terms in the vibration series acting, $D = d_1 + d_3 + d_5 + \dots$

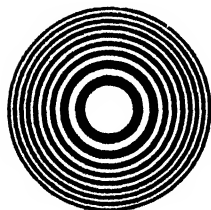


Fig. 98.

This is achieved as follows: A system of black rings of varying breadth is drawn on a sheet of paper. These are then photographed, and the resulting glass negative appears as in Fig. 98. The black zones, if correctly drawn, stop the light from every alternate half-period element, while the transparent zones transmit all those elements which reinforce each other at E . The spacing (assuming a distant source)

is easily obtained from $\rho^2 = nb\lambda$ (equation (2), Article 491), where ρ is the radius of the successive circles that define the half-period zones. For a given wave length and eye position, $b\lambda$ is constant; hence $\rho^2 = nk$, where n is any integer. Therefore the radii of the successive bounding circles must vary as the square roots of successive integers, or $\sqrt{1} : \sqrt{2} : \sqrt{3}$ and so forth. Such a zone plate acts like a converging lens, in that a parallel beam of light falling upon it forms a brilliant focus at a definite distance.

493. Diffraction with cylindrical wave front. Diffraction phenomena are more striking and more easily produced when the source is an illuminated slit, and when other slits or straight edges are used to diffract the resulting waves. If the slit is fairly long, the portion of the wave front which concerns us is cylindrical, with the slit as its axis. The half-period elements are now curved strips bounded by elements of the cylinder. Figure 99 suggests the cylindrical front with horizontal strips bounded by the elements m, m', n, n' , and so forth. The areas of these strips, unlike the half-period zones already discussed, are not equal. They grow smaller at increasing distances

from O . Let a_1 represent the area of the central strip. Let a_2 represent the sum of the areas of the next two strips, mm' and nn' . Let a_3 represent the sum of the areas of the strips $m'm''$ and $n'n''$, and so on. Then by a simple calculation we may show that the ratio $a_2:a_1$ is 0.41, indicating a decrease of nearly 60 per cent. The ratio $a_3:a_2$ is

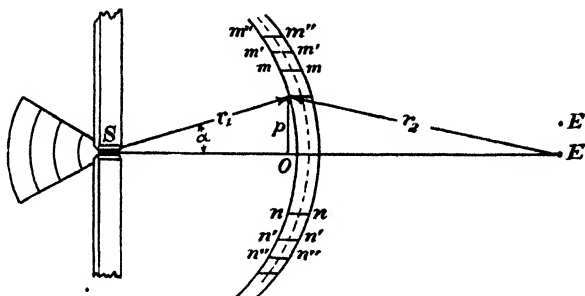


Fig. 99.

0.78, a decrease of only 22 per cent. Thus the areas of successive strips steadily decrease, though the change is more and more gradual as they recede from the pole.

494. Effect of cutting off half-period elements. Suppose a second slit is interposed between the luminous slit S (Fig. 99) and the eye, so that only a limited number of half-period elements passes through. If this auxiliary slit (parallel to S) is gradually opened, the illumination at E steadily increases until, when the entire area a_1 is exposed, the light received at E reaches a maximum of intensity. As the slit is still farther opened, the areas a_2 are exposed. But these, by construction, tend to neutralize the light from the central half-period strip, and the light at E passes through a minimum of intensity. This, however, is far from zero, because a_2 is so much smaller than a_1 . As the slit is still farther opened, a_3 is gradually exposed, and we reach a second maximum due to $a_1 - a_2 + a_3$. This is not so bright as the first, because a_3 is smaller than a_2 and does not fully neutralize its effect of diminishing the light from a_1 . In this way we may pass through a succession of maxima and minima of diminishing contrast. The minima are less and less dark, and the maxima less and less bright, until, when a moderate number of strips have been exposed, there is no more appreciable variation, and the illumination at E reaches a fixed value which is considerably less than when only the central strip is exposed, and is the same as if no obstacle whatever were placed between S and E .

If a fine wire replaces the slit, it blots out precisely those strips which a slit of the same width as the wire would allow to pass. In this case we have to consider the effect of $a_1 - a_2 + a_3 - \dots \pm a_n$,

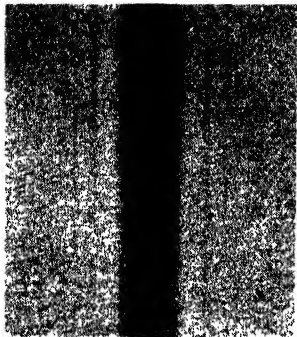


Plate 8.

Photograph of diffraction pattern formed by a fine wire and a slit source, using sodium light. Note the bright line along the axis of the wire, and the "straight edge" bands (Article 496) on either side of the central pattern.

with terms beginning at the left progressively eliminated as we use wires of increasing diameter. This would be impracticable, but we may accomplish the same progressive elimination by moving the eye toward the wire. When E is at the distance which enables the wire just to cover a_1 , the illumination is due to $a_2 + a_3 + \dots + a_n$. This is similar to the original series, which gives unobstructed illumination at E , but with its largest term removed. Therefore the light is considerably diminished. It is still further reduced by moving the eye so near that the wire blots out both a_1 and a_2 , though the decrease is less pronounced than before. Thus as the eye moves toward the wire, the wire covers more and more half-period elements. The illumination steadily decreases, rapidly at first

and then more and more slowly, without undergoing fluctuations of intensity, as when it moves toward a slit of fixed width, or, when at a constant distance from E , the slit is gradually widened.

495. Diffraction pattern due to a narrow slit. We have so far considered only light at a single point E lying on the axis SOE , where the eye perceives varying degrees of illumination according to the number of strips exposed or covered. We shall now consider the distribution of light over a plane surface that includes both E and any other point, as P in Fig. 100. The edges of the slit AB cast geometrical shadows of the slit source S

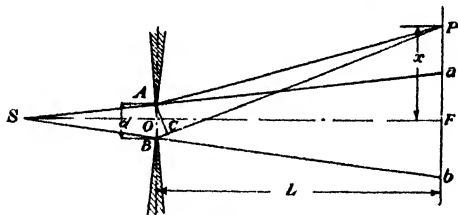


Fig. 100.

above a and below b , while the space between them is illuminated. But actually the point P , when it is inside the geometrical shadow, may be within a region of interference or reinforcement, and so denote

either a dark or bright band parallel to the image ab of the slit AB . If the distance x of P from the axis SOE is such that $BP - AP = BP - CP = BC = \lambda$, then $OP - AP = \lambda/2$, where O lies midway between A and B , and the line AC is drawn so that $AP = CP$. The light at P from the upper and lower halves of the slit evidently differs in phase, point for point, by half a period, and interferes. But if x were such that $BP - AP = 3\lambda/2$, then P is in a bright band, because dividing AB into three parts, instead of into two, as in the diagram, causes two of the three parts to interfere at P as before but the third gives some illumination there. Thus, proceeding into the geometrical shadow from a and b , we encounter a series of alternately light and dark bands.

If the width of the slit could be made less than the wave length of the light used, the distance BC would always be less than λ for any position of P ; therefore complete destructive interference is impossible. Reinforcement at P would be still less possible, for then BC would have to equal $3\lambda/2$. Therefore in this case the illumination grows gradually less without fluctuating, as P is taken farther and farther into the geometrical shadow.

The absence of a definite shadow when the width of a rectangular opening is less than a wave length, applies also to sound.

The width of a door is usually less than the wave length of many tones of the human voice, and as is well known, its frame casts no marked sound shadow. We do not have to see persons through an open door in order to hear them. But the sides of a wide opening, as that between two neighboring houses, cast pronounced sound shadows.

496. Diffraction pattern due to a straight edge. If the upper jaw A of the slit in Fig. 100 is removed, the pattern on the screen is considerably altered. Within the geometrical shadow the illumination falls off gradually, just as it does with a wire of increasing width, because the upper series of half-period strips, Om , mm' , $m'm''$, and so forth, are progressively blotted out by the lower edge B as E moves downward below b and the pole O of the wave front moves downward below B . At b the illumination is due to all the strips above O (now coinciding with B), and is therefore half as bright as if there were no obstacle at all. But if E is above b , O is above B , and we begin to

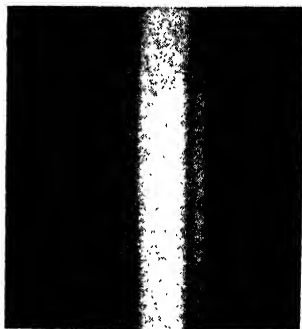


Plate 9.

Photograph of diffraction pattern formed by a narrow slit with slit source, using sodium light.

uncover part of the lower series of strips. If E is so placed that the edge B uncovers the entire area a_1 of the central strip, the illumination is very much brighter than at any other part of the pattern. Still farther up from b , the second strip of the n series is uncovered. This

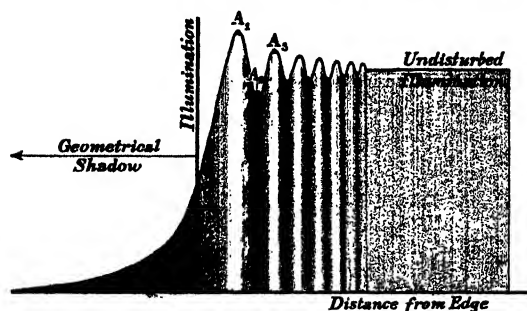


Fig. 101.

tends to reduce the brightness, as now the entire area a_2 interferes destructively with a_1 . Then when a third n strip is exposed, the illumination is somewhat increased. In this way, above the geometrical shadow, we pass through a

succession of maxima and minima of diminishing amplitude, until we arrive at unobstructed illumination. These fluctuations are shown in Fig. 101, where the ordinates of the wavy line indicate relative illumination as suggested by the shading. The Y axis is the knife edge illuminated by a vertical slit, and the geometrical shadow lies to the left. In line with the knife edge, the illumination is exactly half what it would be without obstruction. A little beyond, the illumination reaches a peak at A_1 , which is followed by a minimum at A_2 , and then by a second narrower and fainter maximum at A_3 , and so on. Only a few of these peaks are visible, even under the best conditions with strictly monochromatic light, and the illumination soon reaches the unobstructed value.

497. Diffraction by a rectangular aperture. This is similar to the pattern produced by a narrow slit, except that we now have four edges to consider and therefore two intersecting patterns. If Q in

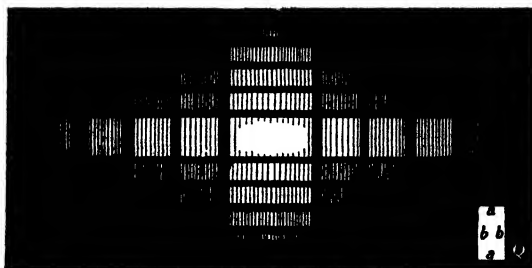


Fig. 102.

Fig. 102 represents the opening, the resulting pattern produced by a point source is seen to be short where Q is long and long where Q is

short. This is to be expected because, in general, the narrower the slit, the wider and more widely spaced are the bands.

The chief pattern is a cross in which the horizontal images are the diffraction bands due to the *bb* edges. The vertical images are due to the more widely separated *aa* edges. In addition to these two series are other much fainter images between the arms of the cross. These also are fully accounted for by the theory.

Such patterns are familiar to anyone who has looked at a bright and concentrated light through the mesh of an umbrella, or other texture in which are small and regular openings. Their number increases the brilliancy of the patterns, which, however, should be the same as with a single opening of the same shape. The orientation of the pattern depends upon the angular position of the edges of the rectangles of which the mesh is composed, so that the cross may be rotated by rotating the texture through which the light is seen.

A circular opening produces circular rings, and a screen perforated with many minute holes close together would form a halo of colored rings about a strong point source. The colors are due to the fact that, for each wave length, there is a distinct set of rings that coincide less and less at greater angular distances from the source. If a transparent screen, sprinkled with minute opaque dots of the same area and arrangement as the pinholes just considered, is interposed between a point source and the eye, the same colored rings appear as with the pinholes. This is explained by the theorem of "complementary screens," which states that at a given point in a diffraction pattern, *where there would be no light without diffraction*, two complementary screens like those just supposed produce the same illumination. The colored rings seen around lights shining through a fog are illustrations of this phenomenon.

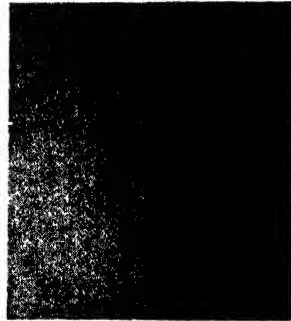


Plate 10.

Photograph of diffraction pattern formed by a straight edge and a slit source, using sodium light.

SUPPLEMENTARY READING

- R. W. Wood, *Physical Optics* (Chap. 7, pp. 218–235), Macmillan, 1934.
 T. Preston, *The Theory of Light* (Chap. 9, Section 1), Fifth Edition, Macmillan, 1928.

PROBLEMS

1. Calculate the radius of a hole in a diaphragm which just uncovers the first half-period element of a plane wave of light whose wave length is 5461 \AA (green line of mercury), when the eye is 3 m in front of the diaphragm.

Ans. 1.28 mm.

2. Show that the opening of the diaphragm of Problem 1 must have a radius approximately $\sqrt{2}$ times as large to uncover two half-period elements, $\sqrt{3}$ times as large to uncover three half-period elements, and so forth.

3. A pinhole of 0.4 mm radius is illuminated with parallel light of wave length $5 \times 10^{-5} \text{ cm}$. How far from the pinhole must the eye be placed to see the first maximum? How far to see the third minimum? *Ans.* 32 cm; $5\frac{1}{3} \text{ cm}$.

4. In Fig. 100, show by means of the nearly similar triangles ABC and OPE that $x/L = BC/d$. If $L = 150 \text{ cm}$ and $d = 0.4 \text{ mm}$, how far from the axis is the first bright band of light of $5 \times 10^{-5} \text{ cm}$ wave length? *Ans.* 2.8 mm.

CHAPTER 38

Fraunhofer Diffraction

498. Two kinds of diffraction. In the illustrations of diffraction we have so far considered, the pattern was formed directly by a plane, or *diverging*, wave front whose path was modified by some kind of obstruction or opening. This type is known as **Fresnel diffraction**.

If a lens is used to *converge* the light so as to form a real image of the luminous source, then the patterns caused by openings or obstructions placed between the source and the converging lens are illustrations of **Fraunhofer diffraction**.

Though not absolutely necessary, in producing Fraunhofer patterns it is desirable that the wave front of the light to be diffracted should be plane, and thus form the diffraction pattern in the focal plane of the lens. The source of light should either be a long way off, or a *collimating* lens should be used, with a slit or point source at its principal focus, the latter forming an "artificial star." As this arrangement greatly simplifies the theory, and is almost always used in practice, it will be assumed in what follows.

499. Narrow rectangular opening. Let us suppose that light from a slit source S in Fig. 103 (a) is made parallel by the collimator L and

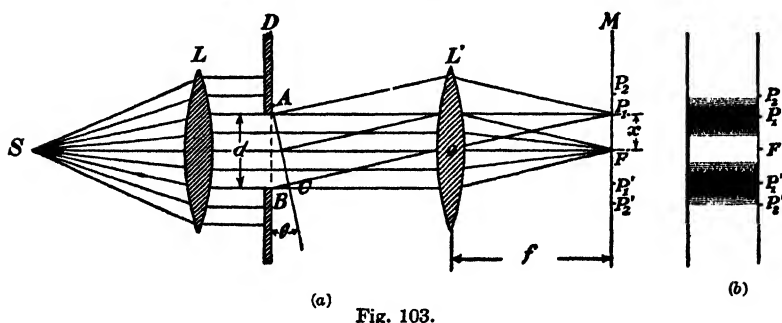


Fig. 103.

is then passed through the narrow rectangular aperture AB , whose width is d . Beyond AB , a lens L' forms a real image of the source S in its focal plane at F . If a screen M is placed there, a bright line image of S is formed on it, as shown in (b), where the screen is repre-

sented as seen from the lens. This central image of the diffraction pattern is always bright in Fraunhofer diffraction, and ordinarily is the only one observed.

It might be thought that as the rays indicated in Fig. 103 reach F by different routes, there could be no certainty that they would reinforce each other there and produce a bright line. They might conceivably arrive in a variety of phases, which would result in partial or complete interference. But this is not the case. It follows from a theorem due to Fermat that the time required for light to go from the source S to its image F is the same for all the rays. Those farther from the axis actually have longer paths, but they have a smaller thickness of glass to go through, and so are less retarded. This decreasing retardation by the glass just compensates for the increased *geometrical* length, and the *optical* paths are said to be the same.

Now consider a point P_1 at a distance x from F . The elementary "Huygens' strips" of the rectangular opening may be taken as sources with respect to P_1 as well as with respect to F . These elements have equal amplitudes, and if the angle θ is small, they produce approximately equal disturbances at P_1 . But in this case there are path differences to be considered. The path from B to P_1 exceeds that from A to P_1 by the distance BC , where AC is normal to the inclined parallel rays. If BC is one wave length, the upper and lower halves of AB destroy each other by interference. Thus P_1 is the center of the first minimum with a similar dark band P'_1 below F . These points are readily located from the similar triangles P_1OF and BAC , whence

$$\sin \theta = \frac{\lambda}{d} = \frac{x}{f}, \text{ very nearly,} \quad (1)$$

or

$$x = f\lambda/d.$$

When $BC = 3\lambda/2$, we may divide AB into three equal strips whose widths are Aa , ab , and bB . Then the path bP_2 is a wave length longer than AP_2 , so that the two upper strips destroy each other at P_2 . This leaves the third strip bB to produce illumination at P_2 , the locus of the next maximum above F . This bright band and P'_2 below F are therefore much fainter than the central band. The distances P_2F , calculated as above, give

$$\sin \theta = \frac{3\lambda}{2d} = \frac{x}{f}, \text{ very nearly,} \quad (2)$$

or

$$x = \frac{3f\lambda}{2d}.$$

The variation of illumination as P recedes from the axis is shown by the solid curve as a function of the distance x in Fig. 104, and the various values of BC in terms of λ are also indicated. The central band is twice as broad as those on either side of it, because it is in a sense the sum of two bands on either side of the axis at F . The intensity of illumination of the first and second diffraction bands may be calculated from the theory, and are found to be $1/22$ and $1/62$ of the illumination of the central image, respectively. Thus they are too faint to be easily observed.

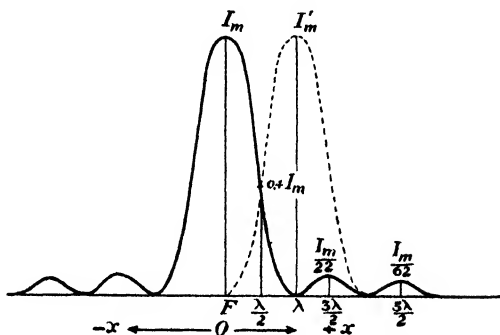


Fig. 104.

500. Rectangular opening and two sources. In this case each source forms its own pattern, similar to Fig. 104. If the angular separation of the sources is small, the two maxima, I_m and I'_m , will overlap to such an extent that the eye cannot see them as separate images. But if I'_m coincides with the first minimum of the pattern of I_m , as indicated by Fig. 104, the two images can be seen as separate, and they are said to be *resolved*. The angular width of each half of the central image is found from $\sin \theta = \lambda/d$ (equation (1), Article 499), or $\theta = \lambda/d$, very nearly. Then the angle α (Fig. 105), which the two images (and therefore the two objects) subtend at the lens, must at least equal θ if they are to be resolved. That is, $\alpha = \lambda/d$ is the test of resolution, and is known as

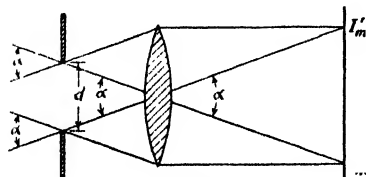


Fig. 105.

501. Resolving power of a lens. If the opening which determines the Fraunhofer diffraction pattern is circular instead of rectangular, a point source forms a circular image with diffraction rings around it, instead of with bands on either side of it. The theory which gives the position of the successive maxima and minima is more complicated than the one for slits, but the result is similar. In this case the criterion of resolution is

$$\alpha = \frac{1.22\lambda}{d}, \quad (1)$$

where d represents the diameter of the opening instead of the width of the slit. If the diaphragm containing the opening is removed, the whole lens is exposed, and its diameter replaces that of the circular opening in equation (1). This is illustrated in Fig. 106, where light

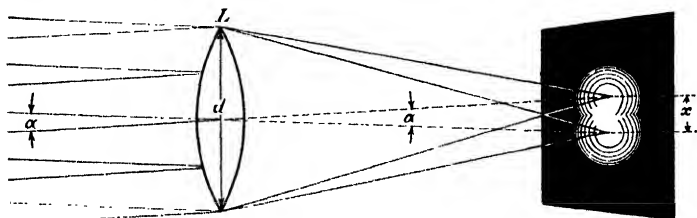


Fig. 106.

from two stars subtends an angle α at the objective L of the telescope. The images of the stars are seen just resolved, with the center of one on the extreme periphery of the other, in accordance with Rayleigh's criterion.

Resolving power of a lens is measured in terms of the minimum possible angular separation of two objects which it can form into distinct images, and this power increases as α decreases. Therefore the resolving power of a lens varies directly as its diameter and inversely as the effective wave length of the light. This wave length may be



(a)



(b)

Plate 11.

(a) Photograph of pinholes illuminated by sodium light and viewed through a third pinhole. Note portions of first and second diffraction rings. (b) Photograph of pinholes so close together that they are imperfectly resolved because of overlapping of their diffraction rings.

either that of the D lines of sodium, which are near the brightest part of the solar spectrum, or a general average such as 5×10^{-5} cm.

If instead of white light, monochromatic violet could be used, the resolving power of a lens would be considerably increased. This idea is employed in the ultraviolet microscope, where invisible light of very

short wave length is used in connection with quartz lenses and a photographic plate to receive the image. Thus the structure of an object otherwise too fine for resolution may be photographed and examined.

502. Calculation of resolving power. Consider a telescope whose objective has a diameter of 20 cm. Then since $\lambda_D = 56 \times 10^{-6}$ cm, $\alpha = 1.22 \times 56 \times 10^{-6}/20 = 3.416 \times 10^{-6}$ radians. But a radian equals 206,265 seconds of arc; therefore $\alpha = 0.7''$ very nearly. This means that two stars whose angular separation is $0.7''$ are barely resolved by the lens, while if the diffraction discs are to just touch each other, the angular separation must be twice as great.

A telescope with a large objective can separate stars that one with a small objective would be quite powerless to resolve, even though both might have the same magnifying power. This is one reason for the use of objectives of large diameter, quite apart from the desirability of collecting a large amount of light from the fainter stars and nebulae which could not otherwise be seen at all.

The resolving power of the human eye is very poor compared to that of a telescope objective; therefore, although the star images referred to above might actually be resolved, they would not be *seen* as such, unless the magnifying power of the telescope were correspondingly great.

It is easy to measure the resolving power of the eye by ruling parallel lines close together on a piece of paper, placing it in a strong light, and observing the lines from increasing distances. When they no longer appear as separate lines, the limiting angle of resolution has been reached, and this angle can be calculated from the known spacing of the lines and their distance from the eye. If then the lines are viewed through a pinhole punched in a card, the observer must come nearer them in order to see them resolved into separate lines.

503. Two rectangular openings. Two narrow slits close together, like a single slit, give rise to a diffraction pattern of the Fraunhofer kind, but in addition they produce an interference pattern such as was described in Article 481. Let light from a narrow slit be made parallel by a collimating lens and fall upon two other narrow slits close together, equally wide, and parallel to the slit source. Such slits are readily improvised by ruling fine cuts with a pen knife in the film of a fogged photographic negative. In Fig. 107 (a), the lens L' forms two real images of the slit S at the same point F . Each is from light passing through one of the slits A and B . The combined image at F has the usual diffraction bands above and below it, as represented

in Fig. 104, but these have twice the amplitude and are four times as bright because there are two slits instead of one.

In addition to the simple pattern just described, there is superimposed upon it an interference pattern of narrower bands, as stated above. These are quite widely spaced when A and B are close together, but become increasingly fine as the distance d between A and B is lengthened. The conditions for the formation of these interference bands are shown in Fig. 107(b). Let P (Fig. 107(a)) be the

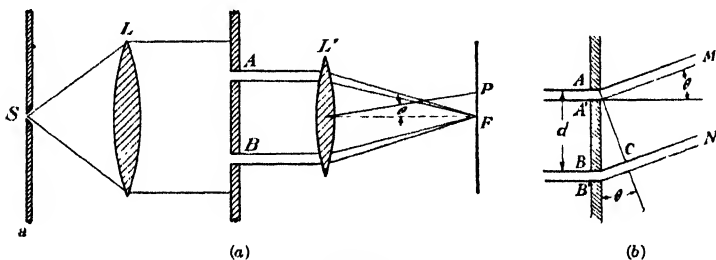


Fig. 107.

locus of a dark interference band whose angular separation from the axis is θ . Then the rays AM and BN must differ in phase by π radians when they arrive at P . This means that if AC is normal to the rays AM , BC must be half a wave length. Thus all the light from the slit AA' will be in opposite phase, point for point, to the light from BB' . If this condition is fulfilled, since BC is normal to AC , $\sin \theta = \lambda/2d$.

In general, dark bands occur when BC is an odd number of half wave lengths, and bright bands occur for an even number of half wave lengths, so that

$$\sin \theta = (2n + 1) \frac{\lambda}{2d} \quad (1)$$

locates a dark band, and

$$\sin \theta = 2n \frac{\lambda}{2d} \quad (2)$$

locates a bright band. The resulting pattern, when both d and the width AA' of the slits are fairly small, is indicated in Fig. 108, where

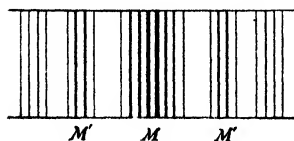


Fig. 108.

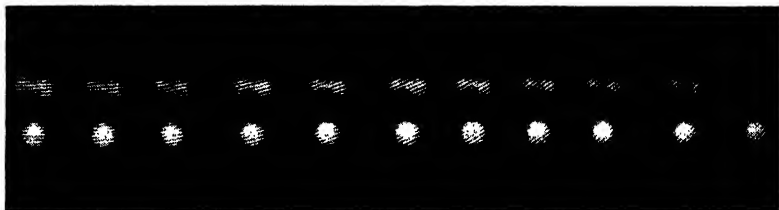
M represents the central diffraction image crossed by eight dark bands. The two adjacent maxima M' are more faint, and of course only about half as wide, as was shown in Fig. 104. The third maxima, in reality barely visible, are about the

same width as the second, and crossed by the same number of dark bands with the same spacing.

504. Astronomical application. If instead of a single source S , there are two sources, like two stars close together, two patterns, like Fig. 108, are formed. If these patterns are coarse and the angular separation α of the stars is very small, then the two patterns coincide so closely that the observer sees only one. But if d is progressively increased, the patterns become finer and finer and then their small angular separation begins to be apparent, just as two pieces of cheesecloth held one over another against a window pane look quite different according to whether the meshes coincide or are "out of step."

Suppose then that the distance between the two slits is slowly increased until the two very fine patterns are out of step, so that the dark bands of one coincide with the bright bands of the other. Then the bands disappear and the image M of the double star (Fig. 108) looks like that of a single star if its "components" are equally bright. If they are not, the bands pass through a minimum of intensity but do not disappear. The pattern crossing M is restored if d is further increased until the two systems are once more in step.

In the manner just described, with two slits over the objective of a telescope, it is possible to measure the angular separation of the components of a double star that are too close together to be resolved by



Courtesy of Carnegie Institution of Washington.

Plate 12.

Interference patterns from artificial stars, taken in the Pasadena laboratory of Mt. Wilson Observatory.

any telescope. The distance d between the slits is increased until the clearness of the interference pattern reaches a minimum. Then the angle subtended by the two stars is given by $\alpha = k\lambda/d$, where λ is the effective wave length of the star's light, and k is a constant whose value (1.22 to about 1.45) depends upon the character of the star observed. If the distance of the double star from the earth is also known, the linear separation of its components may be calculated.

One of the most brilliant achievements of modern astrophysics is the measurement of the angular diameters of some of the brighter stars such as Betelgeuse, Antares, Aldebaran, and so forth. This

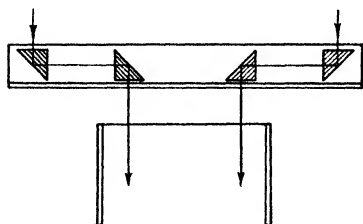
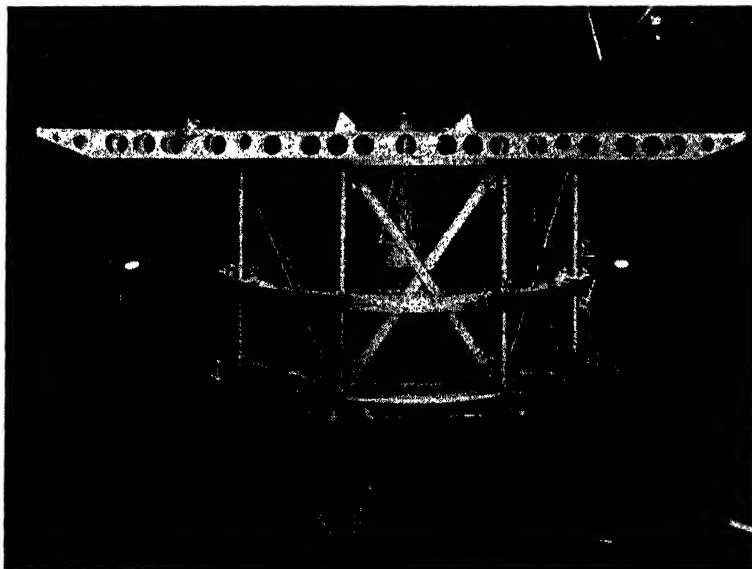


Fig. 109.

method, devised principally by Michelson, is like that just described, but here the two opposite sides of the star's disc take the place of the two components of the doublet. As α , the angle subtended by the star at the telescope, is so excessively small, the distance d must be much greater than the diameter of the objective of any

existing telescope. In order to provide this necessary distance between the slits, a "beam interferometer" 20 feet long spans the top of the great reflector at Mt. Wilson. On this beam are mounted two pairs of relatively narrow prism mirrors. The outer pair corresponds to the two slits, and is movable up to a total separation of 20 feet, each prism being 10 feet from the axis of the telescope (Plate 13). These reflect



Courtesy of Carnegie Institution of Washington.

Plate 13.

Twenty-foot beam interferometer mounted on 100-inch reflector at Mt. Wilson, showing outer prisms 12 feet apart,

the light to the pair of fixed-prism mirrors directly over the telescope, where the light is bent a second time through 90° and enters the telescope parallel to its axis, as indicated diagrammatically in Fig. 109.

505. The diffraction grating. There are two kinds of gratings—transmitting and reflecting. Reflection gratings are used in most important investigations. They are ruled with a diamond on speculum metal. The ruled lines scatter the light so that it is regularly reflected only from the spaces between. These when illuminated become the source of Huygens' wavelets as if they were so many parallel slits. But since such gratings are very expensive, those most commonly used for demonstration and simple experiments are collodian "replicas" of the original gratings. The collodian film has ridges corresponding to the ruled lines. These also scatter the light, so that the smooth spaces between act like slits.

As the theory of the transmission grating is the simpler one, it alone will be considered here. We may then imagine a grating composed of

a system of parallel slits close together, which is shown greatly enlarged at G in Fig. 110. Each slit, when illuminated, becomes the source of Huygens' wavelets. If the individual slit is very narrow, the light is diffracted through quite a large angle θ away from

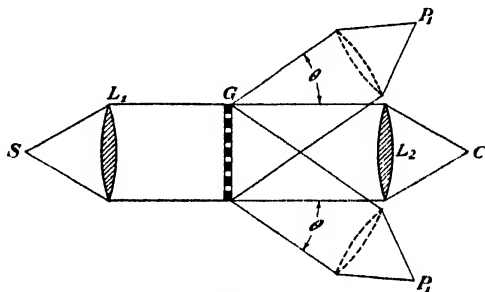


Fig. 110.

the axis of the incident beam, just as if there were only one slit. In the diagram, a slit S is shown illuminated by a monochromatic source not shown. The light from S , made parallel by the lens L_1 , passes through the grating G and is concentrated in the central image C by the lens L_2 . Although the geometrical distances from the various slits to C are different, the optical paths are the same, in accordance with Fermat's principle. Therefore light rays from all the slits arrive at C in the same phase and reinforce each other there without regard to the wave length of the source.

At points such as P_1 , there may be reinforcement, provided the light from the successive slits differs in phase by an integral number of wave lengths. The conditions which determine this reinforcement are readily understood from Fig. 111, where the construction is identical with Fig. 107 (b), except that now there are more slits. As before,

$AB''C''$ is drawn normal to the diffracted rays. Then if $BB'' = \lambda$, $CC'' = 2\lambda$, $DD'' = 3\lambda$, and so forth, the light from each slit reinforcing that from every other slit at P_1 , since the path from each slit is one wave length longer than the path from the slit just above it, point for point.

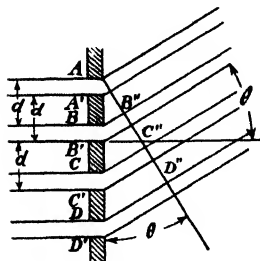


Fig. 111.

The condition of reinforcement at P_1 is fulfilled when the distances like AB (or $A'B'$), known as the "grating element," d , satisfy the equation $d \sin \theta = \lambda$. But we should also have reinforcement if $d \sin \theta = 2\lambda$ or 3λ , or any integral multiple of the wave length. Therefore, in general, $\sin \theta = n\lambda/d$, where n is any integer. We may also set

$N = 1/d$, where N is the number of ruled lines per linear centimeter, and is called the grating constant. Then the grating equation becomes

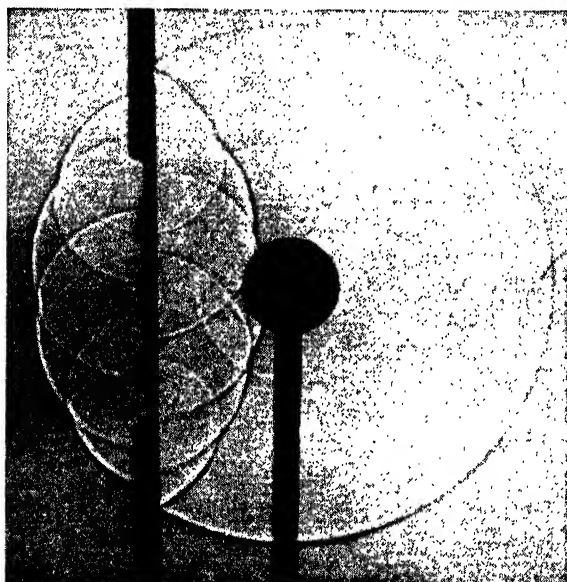
$$\sin \theta = nN\lambda. \quad (1)$$

When $n = 0$, we have the central undiffracted image. If $n = 1$, the points P_1 indicate the *first-order spectra* for light of wave length λ . These are located by $\sin \theta_1 = N\lambda$. When $n = 2$, the spectra are farther from the axis at points P_2 , defined by $\sin \theta_2 = 2N\lambda$. Third-order spectra are located by $\sin \theta_3 = 3N\lambda$, and so forth.

Spectra of increasingly higher orders lie farther and farther from the axis. Their total number is limited by the narrowness of the slits of the grating, for the spectra are clearly visible only within the region of the central diffraction image such as would be formed by any one of the slits. Then if the slits are narrow but relatively far apart, a large number of orders very close together are visible. If the spacing is finer, the spectra are more widely spaced, and consequently fewer are visible. If the slits are wide, very few are visible in any case.

506. Grating spectra. If white light falls upon the slit S , the many wave lengths present are spread out in a spectrum. The violet light, having the shortest wave length, is bent least, while red, having nearly twice as great a wave length, is bent through nearly twice as large an angle. From this it is apparent that the second-order spectrum begins with violet just beyond the end of the red of the first order, because $2\lambda_V$ is a little larger than λ_R . But the third-order violet falls within the second-order spectrum, because $3\lambda_V$ is less than $2\lambda_R$. In fact, this results in an overlapping beyond the yellow-green, and renders half of the second-order grating spectrum of white light almost useless.

In order to have a spectrum in which the colors are pure, and the absorption or emission lines sharp, it is essential that the light received at a given point, as P_1 (Fig. 110), shall be practically monochromatic. In the elementary theory given in the last article, this question was not raised, and in fact, with the few openings shown,



Courtesy Professor A. L. Foley, Indiana University.

Plate 14.

Photograph of sound wave meeting a plane grating. Wavelets due both to transmitted and reflected sound illustrate how a diverging beam of light behaves with a grating of much finer ruling.

monochromatic light from the source, though brightest at P_1 , would by no means be extinguished at other nearby points. It would gradually fade out with a change of θ , until the path difference from adjacent openings became equal to $\lambda/2$ or $3\lambda/2$, when interference would be complete for that wave length. But this is not the case when there are many slits.

Let us now suppose that we are dealing with a grating having 10,000 ruled lines (or spaces) and illuminated with white light. Light whose wave length has a particular value λ_1 is seen at an angle θ which satisfies the equation $\sin \theta_1 = N\lambda_1$, and forms a first-order image of the slit S . Now consider a wave length λ_2 which differs from λ_1 by one part in 10,000, or .01 per cent. Then the first few openings send out

beams in the direction θ practically in phase with each other, and there is no appreciable interference. But the phase difference between them steadily increases, each succeeding space adding .01 per cent to its value, until at the 5001st space, this amounts to $5000 \times .01$ per cent, or 50 per cent phase difference. In other words, light of wave length λ_2 , coming from the 5001st space, is in opposite phase from that from the first space, and they destroy each other completely at the point where the light of wave length λ_1 was perfectly reinforced. The second space and the 5002nd interfere in the same way; the 5003rd destroys the third, and so on, so that all the light from the first 5000 spaces is completely destroyed by light from the second 5000.

We have thus shown that only light differing in wave length by less than .01 per cent from λ_1 can be seen at P_1 . This means that the lines in an absorption or emission spectrum formed by the grating are fairly narrow and sharp, and that the colors in a continuous spectrum are fairly pure. This result was achieved by having many grating spaces, regardless of the fineness of the ruling. So it is evident that increasing the *total number* of ruled lines increases the fineness of spectral lines and the purity of spectral colors. On the other hand, closeness of ruling indicated by a small value of d , or a large N , results in longer spectra with increased resolution of their parts.

The following conclusions regarding the grating may now be made: The sine of the angular deviation is directly proportional to the wave length, which is not true of prism spectra; therefore if N (or d) is known, the wave length of a given spectral line is easily measured. The angular deviation for a given wave length, and consequently the angular dispersion of the whole spectrum, depend upon N . Purity of the spectrum depends upon the *total* number of ruled lines, and is therefore enhanced for a given value of N by having a long grating. The grating spectrum is much less brilliant than the prism spectrum because the light of a given wave length forms not only part of the central white image at C where all waves come together, but also part of the two spectra for each of the orders represented, whereas in a prism spectrum a given wave length has only one path. The number n of orders of the spectrum that may be formed by a grating of a given width of slit decreases as the value of N increases, so that if N is very large, only the first and part of the second order are produced, but if N is small, many orders with correspondingly small dispersion may be found.

SUPPLEMENTARY READING

T. Preston, *The Theory of Light* (Chap. 9, Section 2), Fifth Edition, Macmillan, 1928.

PROBLEMS

1. A lens whose focal length is 40 cm forms a Fraunhofer diffraction pattern of a slit of 0.3 mm width. Calculate the distance from the axis of the first dark band and of the next bright band, using sodium light. (Mean $\lambda = 5893 \text{ \AA}$.) *Ans.* 0.79 mm, 1.18 mm.

2. What is the angle of resolution (α) of the slit in Problem 1, using light of the same wave length? *Ans.* $6\frac{3}{4}'$.

3. Calculate the angle of resolution of a lens whose diameter is 8 cm, using sodium light. *Ans.* $1.85''$.

4. Two narrow slits are 0.6 mm apart. Using light of 5×10^{-5} cm wave length, calculate the angle which locates the third dark band ($n = 2$). *Ans.* $7' 9''$.

5. Calculate the distance between dark bands in the pattern formed, as in Problem 4, by a lens whose focal length is 80 cm. *Ans.* 0.66 mm.

6. A transmission grating has 4000 lines/cm. Calculate the angular deviation of the second-order spectrum of sodium light. *Ans.* $28^{\circ}8'$.

7. A grating ruled with 6000 lines/cm forms the first-order spectrum of a certain line at an angle of 18° . What is the line's wave length? *Ans.* 5150 \AA .

8. Calculate the grating constant (N) of a grating which forms the third-order spectrum of sodium light at 1.6 cm from the central image on a screen 80 cm from the grating. *Ans.* 113 lines/cm.

CHAPTER 39

Polarized Light

507. Meaning of polarization. In general, the word **polarization**, or **polarized**, means that a body or a phenomenon takes on a changing aspect according to the direction from which it is considered. Thus the motion of the earth around its axis appears quite different according to whether it is viewed from one of its poles or elsewhere. A magnet also has poles, and in consequence its properties are different in different directions. A beam of light would seem to have no such preferred points or planes, and in general this is the case, because the behavior of a reflected or refracted beam of ordinary light is quite independent of the plane of incidence, which contains both the incident and the reflected or refracted ray.

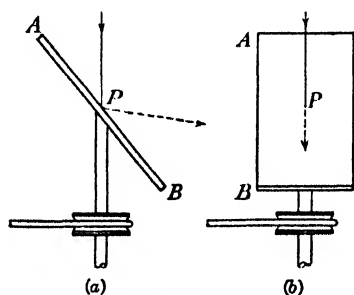


Fig. 112.

Thus if a ray of light is incident on a mirror AB , as shown in Fig. 112 (a), the ray reflected at P , indicated by the dotted line, together with the incident ray, determines the plane of incidence, which is that of the paper. But if the mirror is rotated through 90° about an axis, as indicated, then (b) shows that although the angle of incidence is the same as before, the

incident ray now comes toward the observer, and the plane of incidence is normal to the paper. Such a rotation of the mirror, however, does not alter either the angle of reflection or the intensity of the reflected beam, which is constant even when the mirror is rapidly rotated. We may then conclude that a beam of ordinary light is the same on all sides, since there seems to be no plane of incidence which favors reflection more than any other. If there were such a plane, it would mean that if we could look at a beam of light from the side, it would have different aspects according to our viewpoint, just as a knife blade looks different when seen sideways or edgewise.

508. Plane polarization. The lack of a preferred plane, indicated by the experiment just described, seems inconsistent with the picture

of light propagation already given, in which it is regarded as a transverse vibration. The direction of vibration must have a definite orientation, and apparently it should have different properties according to whether light is thought of as vibrating at right angles to the observer, as it streams past him, or edgeways, like the knife blade. But if there is such a direction, and if it changes nearly a billion times per second, as seems likely (see Article 480), then it may be thought of as assuming all possible orientations about the ray as an axis, in an incredibly short time, and as all directions are equally represented, none of them is preferred.

But now suppose the vibrations are limited to a single plane; then the beam is polarized, and we should look for a difference in its behavior under the varying conditions of reflection brought about by rotating the mirror described above. Such a beam can be produced, as we shall see, and is then said to be **plane polarized**. Other forms of polarization are possible, such as elliptical or circular, when the vibrations are in elliptical or circular paths about a ray as an axis. But at present we shall consider only the simpler case of plane polarized light.

509. Polarization by reflection. When light is reflected from a nonconductor of electricity, such as water, glass, varnish, and so forth, it is found to be partially plane polarized. This was discovered by Malus in 1808. By "partial polarization" is meant that, on the whole, the waves vibrate more in one plane than in any other, as suggested in Fig. 113, where the vectors representing the displacement are shown lying in various planes about a single ray viewed end on. These planes show a preference for the plane whose trace is PQ , where it intersects the paper, and the group is said to be partially polarized in the plane whose trace is MN for a reason about to be explained.

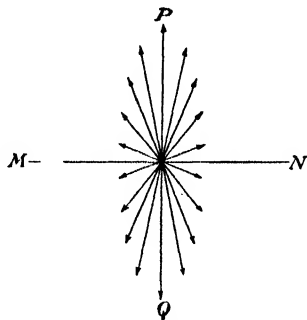


Fig. 113.

In Fig. 114, two panes of glass are so related that the light incident at 45° on P , called the *polarizer*, is partly transmitted, and partly reflected to A , the *analyzer*, where again it is partly reflected and partly transmitted. In (a) the two panes are parallel, and the beam r , after the second reflection, is found to be stronger than the transmitted beam t . In (b) the lower pane has been turned through 180° about a vertical axis, such as was shown in Fig. 112, and the relative

intensities of r and t are found to be the same as in case (a). But when the lower pane is rotated through 90° , as shown in (c), the re-

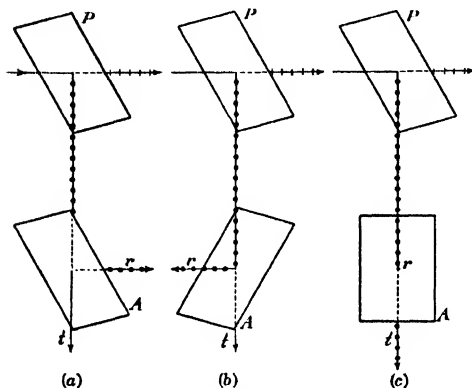


Fig. 114.

flected ray r , normal to the paper, is less intense than the transmitted ray t . This shows that the light was partially polarized by the first reflection, and the plane of polarization has been taken as the plane of incidence, which is the plane of the paper. It is, however, quite certain that the plane of the vibrations of a polarized ray of light is per-

pendicular to the plane of polarization. This is like saying that the earth is polarized in the plane of its equator, because the poles lie on opposite sides of this plane. Although such a definition is logical, it is clearer to explain the phenomena of polarized light in terms of the planes of vibration. These planes are indicated by dots when the vibrations are in a plane normal to the paper, and are indicated by short transverse lines when the vibrations are in the plane of the paper. Thus the ray transmitted by P , in Fig. 114, is partly polarized to vibrate in the plane of the paper, while the reflected ray vibrates in a plane normal to the plane of the paper.

Referring again to our panes of glass, we may make an examination of the beam *transmitted* by the first pane, as shown in Fig. 115, where the second pane is free to rotate around the incident ray as an axis. The beam transmitted by the analyzer is stronger than that reflected in both (a) and (b), for their planes of vibration are similarly related to the surface of the glass in each position. But in

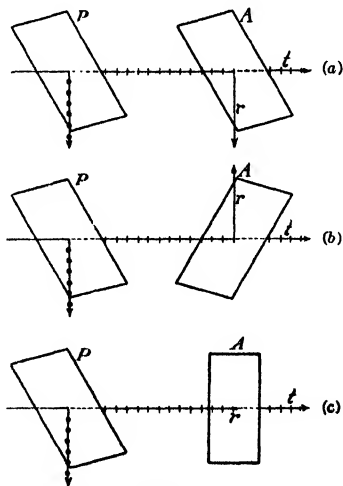


Fig. 115.

(c), when the analyzer had been rotated through 90° , the reflected ray coming directly toward the observer is the stronger of the two, because the vibrations of the incident light are parallel to the reflecting surface, as in Fig. 114 (a) and (b), and this is evidently the condition which especially favors reflection.

510. Brewster's law. In the preceding diagrams the angle of incidence was taken as 45° for the sake of simplicity of construction. But if other angles are used, it is found that the degree of polarization, and consequent differences in the intensity of the beams separated by the analyzer, depend upon the angle. This fact was discovered in 1815 by Sir David Brewster, who formulated the law named for him, that polarization by reflection is most complete when the angle between the reflected and refracted rays is 90° . In Fig. 116, a beam of light is partly reflected and partly refracted at the surface of a medium like glass or water. The resulting beams are partly polarized, as indicated by the dots and cross lines. It is evident that if the angle between Pa and Pb is 90° , the angle of refraction r must be the complement of the angle of incidence i . Therefore $\sin r = \sin (90^\circ - i) = \cos i$, and the index of refraction n equals $\sin i / \sin r = \sin i / \cos i = \tan i$, so that the most effective angle for polarizing by reflection is one whose tangent is equal to the index of refraction of the reflecting medium. The refracted beam is also more polarized at this angle than at any other. Hence the angle which produces the most complete polarization by reflection or refraction is known as the **polarizing angle** of the medium. It is 57° to 58° for ordinary glass, and $53^\circ 7'$ for water at 0°C . A few substances (notably glycerine), whose index of refraction is approximately 1.46, were found by Jamin to produce perfectly plane polarized light by reflection at the polarizing angle. With most substances the reflected light is more or less elliptically polarized and cannot be completely extinguished by another mirror.†

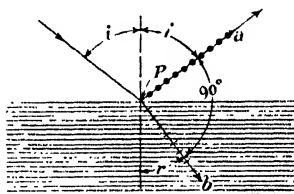


Fig. 116.

511. Double refraction. In discussing the simple phenomena and laws of refraction as described in Chapter 32, it was assumed that the refracting medium was isotropic. That is, it had the same optical properties in all directions because its atomic or molecular structure conformed to no definite pattern. But crystals have a very definite structure due to the arrangement of the atoms in what is

† Paul Drude, *Theory of Optics* (p. 294), Longmans, Green & Co., 1929.

known as a **space lattice**, to be considered more fully in connection with X-ray spectroscopy. Such a medium is **anisotropic**, and the behavior of light is affected by the direction it takes through the crystal.

About the year 1670, Erasmus Bartholinus, a Danish philosopher, discovered that crystals of Iceland spar (calcium carbonate, or calcite) split up a ray of light into two rays, so that objects seen through them appeared double. The meaning of this was not understood at the time, although Huygens was able to explain it in part from the point of view of his wave theory, and discovered the fact that the refracted light was polarized. But it was not until 1808, when Malus discovered polarization by reflection, that general interest in these phenomena was awakened and numerous investigators began to examine the propagation of light in crystals.

All mineral crystals, except those belonging to the cubical system, exhibit double refraction, as well as other substances whose atomic structure conforms to certain patterns. Even isotropic bodies, such as glass, may exhibit double refraction when subjected to strain.

If a black dot on a piece of white paper is viewed through one of the surfaces of a rhomb of Iceland spar, it appears as two dots, each of a paler tint than the original, and if the rhomb is rotated about a vertical axis under the eye looking straight down, as in Fig. 117, then one of the images appears to rotate about the other, which remains fixed. Moreover the fixed image appears a little nearer to the eye.

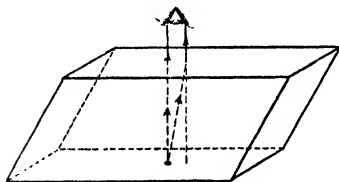


Fig. 117.

Now it is evident that if the rhomb were of glass, the dot would appear single and would not rotate; therefore the immovable image seen through the spar is the normal or *ordinary* image, and the other is the abnormal, or *extraordinary* image. The ordinary image is the one which seems nearer, and this fact shows that Iceland spar has a higher index of refraction for the ordinary ray than for the extraordinary, since, as we have seen, a glass slab seems to bring objects nearer, and the higher its index, the nearer they appear. Thus, from this simple experiment, we know that one of the two beams produced by the double refraction of calcite behaves like ordinary light, and is more retarded than the other, which behaves quite differently.

512. Wave-surface construction of double refraction. Huygens realized that this phenomenon involved two types of propagation, and showed correctly that whereas the ordinary beam had a spherical wave

front in doubly refracting media, the extraordinary beam had an ellipsoidal wave front. Before describing his construction, however, it will be well to explain the meaning of an **optic axis**. This is any direction in a crystal in which both ordinary and extraordinary rays travel with the same speed, and so have the same index of refraction. Some crystals have several such axes, but two of the most important, Iceland spar and quartz, have only one, and we shall here consider only this kind.

If, then, the two rays travel with the same speed in a given direction through a crystal, the line so indicated is an optic axis, and any line parallel to it is an optic axis also. Then, having determined this direction, we can imagine the crystal as being made up of a bundle of an infinite number of such lines all pointing the same way.

In Fig. 118, let MN represent the surface of a crystal of Iceland spar whose optic axis has the direction AX in the plane of the paper.

Let a beam of light bounded by the rays a and b and having a plane front AB be incident on the surface at an angle i . Then with A as a center, draw a circle whose radius is to BC as the velocity of the ordinary ray in the crystal is to the velocity of light in air, and finally draw Cp tangent to the circle. We have now completed

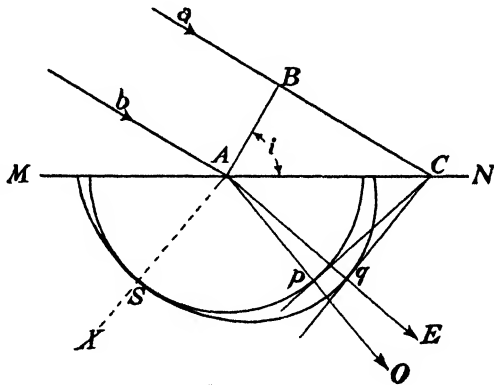


Fig. 118.

Huygens' construction for ordinary light, as in Article 328. The circle represents the spherical wavelet emanating from A , and pC is the wave front of the refracted beam, while the radius Ap is one of its rays.

But there is another mode of propagation which travels out from A as an ellipsoid of revolution. It is tangent to the circle at S on the optic axis, where the two velocities are equal, and its major axis is at right angles to AX . If the ellipse is rotated around the optic axis AX , it develops a surface of revolution that is the wave front of this disturbance emanating from A . A line from C , tangent to it at q , is the extraordinary wave front of the original beam after extraordinary refraction. The line Aq , drawn through the point of tangency,

is a ray, but it is not in general perpendicular to the wave front, as is always the case in isotropic media.

The indices of refraction of the two rays *O* and *E* are obviously different. The ordinary ray is the more bent, and has the higher index because it is more retarded, except along the optic axis. This index is constant for all angles of incidence, according to Snell's law, but the extraordinary ray has a variable index depending upon the angle of incidence and the direction of the optic axis.

Those crystals in which the extraordinary ray travels in general faster than the ordinary, and so has a smaller index of refraction than the ordinary index, are called **negative crystals**. Crystals in which the ordinary rays travel faster, and have the smaller index, are known as **positive crystals**.

Quartz is the most familiar example of the positive crystal. Its wave front construction diagram differs from the preceding in that

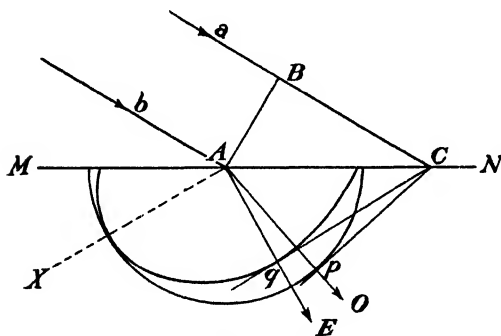


Fig. 119.

the ellipsoid of revolution lies within the sphere, and is tangent to it along its major instead of its minor axis. This is shown in Fig. 119, where the extraordinary ray *E* is the more bent. The ellipsoid of revolution, which is the wave front of the disturbance from *A*, shows that the ex-

traordinary ray travels as fast as the ordinary along the optic axis, but more slowly in other directions, and the two differ most in speed along the minor axis of the ellipse, whereas the greatest difference in negative crystals is along the major axis.

513. Polarization in double refraction. Shortly after the discovery of double refraction by Bartholinus, Huygens, in studying the phenomenon, found that the rays were polarized. This fact is made evident by placing one doubly refracting crystal over another of equal thickness and looking at a black spot on white paper through both. If they are symmetrically placed so that the optic axes are parallel, the second crystal acts simply as if the first had been doubled in thickness, and the two images are more widely separated than before. But as it is rotated about a vertical axis while the eye looks directly down-

ward, four spots appear, and at an angle of 45° between the crystals, these are all equally intense. At 90° there are two spots again, at 135° four, and at 180° only one. After this the same succession of events takes place until the original position is recovered.

The explanation is that the two beams produced by the first crystal are completely polarized in planes at right angles to each other. The ordinary ray is polarized in a plane known as a principal section. In the type of crystals we are considering, this is a plane which includes the optic axis and a normal to the refracting surface. In Iceland spar the optic axis may be determined with reference to a corner, such as

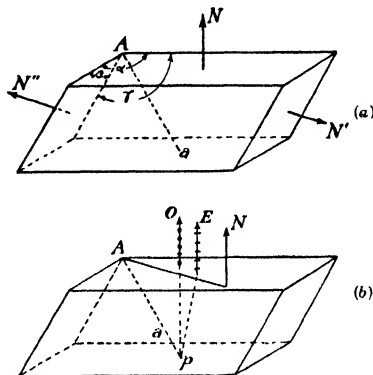


Fig. 120.

A in Fig. 120 (a), where the three angles α , β , and γ , made by the faces meeting there, are all obtuse. Then a line *Aa* which makes equal angles with the three faces or edges is an optic axis, as are all lines parallel to it. Any plane parallel to *Aa* and to a normal *N*, *N'*, *N''*, and so forth, is a principal section. Therefore in Fig. 120 (b) the principal section is defined by *N* and a line drawn from its base parallel to *Aa*, because the upper and lower surfaces to which *N* is normal are the refracting surfaces for the point *p*. If this plane is supposed parallel to the plane of the paper, the ordinary ray is polarized in that plane, and its vibrations are normal to it as indicated in the diagram, while the vibrations of the extraordinary ray are in the principal section.

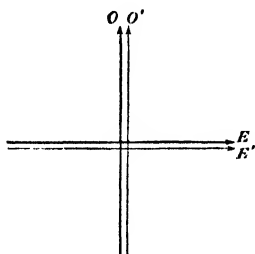


Fig. 121.

514. Vector diagram of polarized beams.

The explanation of the various images seen when one crystal is rotated above another is most clearly understood by the use of vector diagrams in which the direction of the vector represents the plane of vibration when looking at the beam end on, and its length is proportional to the amplitude of the vibration.

Thus in Fig. 121, the two vectors *O* and *E* represent the amplitude and planes of the vibrations of the ordinary and extraordinary rays seen looking down on the crystal, as shown in Fig. 120 (b). Then if another crystal is placed over it, with their princi-

pal sections parallel, the second crystal is, so to speak, in phase with the first, and its effect is to increase the divergence of the rays due to its vectors O' and E' parallel to O and E .

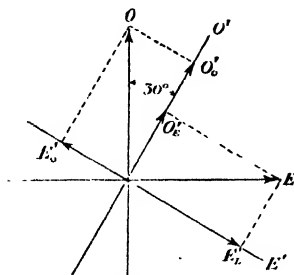


Fig. 122.

But now if the upper crystal is rotated through say 30° , as in Fig. 122, the ordinary ray from the lower one is resolved into two components vibrating in the directions O' and E' . These give rise to a rather strong beam $O'O'$, and a weaker one $E'O'$, while the extraordinary beam is also resolved into two of similar amplitudes, that is, $E'E'$ and $O'E'$. These four evidently become equally intense at 45° . At 90° , O' coincides with E , and E' with O , and only two spots are visible. The other cases are similarly explained until we reach 180° , when the fact that there is only one spot visible is accounted for by the reverse bending of the extraordinary ray, as shown in Fig. 123 (b). This counteracts the separation shown in (a).

515. Nicol's prism.

The most effective method of producing a beam of plane polarized light was invented by William Nicol, a Scottish physicist, after whom the device is named. Nicol's prism is made of Iceland spar, and delivers half of the incident

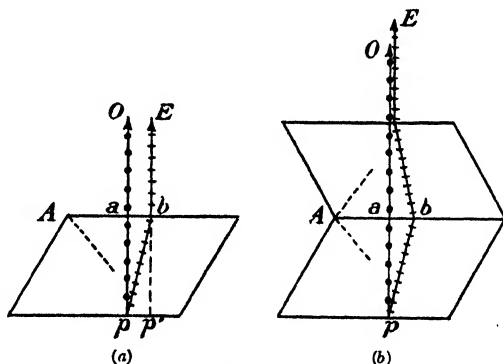


Fig. 123.

light in a completely polarized beam. A rhomb of this substance about three times as long as it is thick is obtained by cleavage from the crystal, which may have a variety of forms. This rhomb is then cut across the ends so as to change the natural angles at C and D from 72° to 68° . The crystal is then cut in two, along a plane perpendicular both to a principal section and to the new surfaces CB and AD , as indicated in Fig. 124, where we may assume the plane of the paper to be a principal section with respect to a beam R . The two parts of the crystal are then cemented together with Canada balsam, whose refrac-

tive index n_c lies between the indices of the ordinary and extraordinary rays, that is, $n_o > n_c > n_e$. When the ray R enters the crystal, it splits up into two rays, as shown. The ordinary ray is the most deviated and is polarized in the principal section, as indicated by the dots.

When it reaches the Canada balsam, it is going from an optically denser into an optically rarer medium, and is totally reflected, because the angle α is greater than the critical angle, which is 59° . The extraordinary ray, however,

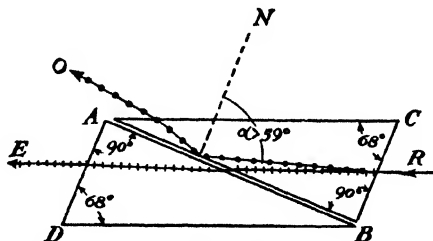


Fig. 124.

in entering the balsam, is going into an optically denser medium, because $n_c > n_e$, and there is no appreciable internal reflection. In passing again into the crystal from the balsam, it is going from dense to rare, and we might expect total reflection, but the angle of incidence is a little less than the critical angle, which is about 57° . This circumstance, coupled with the fact that its plane of vibration is better suited to transmission than reflection, as explained in Article 509, results in an almost complete transmission of the extraordinary beam E . The plane of vibration, as we have seen, is parallel to the principal section, so that the emergent light, when viewed from the end of the crystal, is vibrating in the plane indicated by the diagonal line across the small circle in

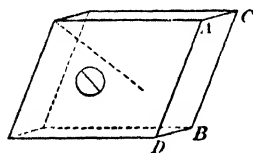


Fig. 125.

Fig. 125, which is parallel to the shorter diagonal of the end face of the crystal.

A pencil of light polarized by a Nicol's prism passes perfectly freely through another Nicol's prism having its corresponding planes parallel to the first. But if the latter is rotated through 90° about the pencil as

an axis, the light is completely extinguished, though it is completely restored at 180° from the original position. Between these positions more or less light passes the second prism according to the angle between them, and the intensity of the transmitted light is found by projecting the vector representing the vibration produced by the polarizer (first prism) upon that of the analyzer (second prism), exactly as E is projected on E' in Fig. 122. Then the intensity

of the transmitted light is proportional to the square of this component.

516. Polaroid. A new material for polarizing light, called *Polaroid* invented by E. H. Land of Boston, was first offered commercially in 1935. It promises to replace the Nicol's prism in many of its applications, as well as to open up new uses because of its relative cheapness and the possibility of greater size.

Polaroid is a film of cellulose acetate somewhat similar to cellophane, but it has imbedded in it, as in a suspension, many minute, synthetic, doubly refracting crystals. These crystals are all oriented the same way by giving the film a stretch in one direction during the process of manufacture. As in the case of tourmaline, each crystal transmits only one beam of polarized light, the other being absorbed within the crystal. In the visible spectrum, about 40 per cent of the light is transmitted instead of an ideal 50 per cent, and the polarization is very slightly imperfect at the two ends of the spectrum.

Films of Polaroid can be produced having far greater area than the sections of the largest Nicol's prisms, and are available for such applications as stereoscopic projection, and the elimination of glare from automobile headlights. The projection involves in principle two lanterns throwing on the same screen images of a pair of films taken from different positions, as in stereoscopic views. By means of Polaroid filters, one lantern polarizes the light vertically, the other horizontally. When seen with the unaided eye, the pictures make a confused jumble, but each person in an audience can be equipped with Polaroid spectacles so oriented that the right eye sees only the right-hand picture, and the left eye sees only the left-hand picture. The result is stereoscopic vision such as has previously been achieved somewhat imperfectly by projecting red and green pictures which are viewed through red and green glasses.

To eliminate headlight glare, both headlights and windshields of automobiles might be equipped with Polaroid having the axis of polarization at 45° with the horizontal. If this angle were in the same "sense" in all cars, then the axes of the light from approaching cars would be crossed at 90° so that windshields would extinguish the light of headlights. They would, however, transmit road illumination as seen through the windshield of the car which produced it.

517. Polarization by scattering. When light passes through a medium in which particles are suspended, its path becomes visible from the side, as in a barn filled with hay dust, where beams of light are strongly marked by the "motes" in their path. This is due to re-

flection from the surfaces of these relatively large particles. But if the suspended particles are so small that their diameters are comparable to the wave length of the light, then the light is not reflected, but diffracted, as when it bends around a fine wire, and each particle becomes, as it were, a secondary source. The light diffracted in this way is partially polarized in the plane that contains both the original and diffracted rays. This fact may be determined by examining both the light *reflected* at the polarizing angle from the surface of a liquid, and the light diffracted from the particles it holds in suspension, as those used by Perrin in studying the Brownian movements.

If a jar, shown in Fig. 126, contains a liquid with fine particles held in suspension (soapy water will answer), we may examine a reflected beam *A*, or a diffracted beam *B*, with the aid of a Nicol's prism, as indicated. It will then be observed that if the prism is set to extinguish *A* as much as possible, it will also extinguish *B* as much as possible, without rotating it about its axis. This is strong evidence for the view, mentioned in Article 509, that the direction of vibration, indicated by the heavy dots, is normal to the plane of polarization defined by *D* and *A*, the incident and reflected rays. For if this were not the case, it would mean that the vibrations of the vertical beam *D*, which are all normal to its direction, as indicated by the radiating arrows, must have been rotated through 90° by the suspended particles, and would then be represented by the dotted cross-lines of ray *B*. It is much more plausible to assume that the vibrations are correctly indicated by the heavy dots, which represent polarization due to a selective effect without any rotation. Therefore, since both *A* and *B* are said to be polarized in the plane defined by *D* and *A*, it follows that their vibrations must be normal to the so-called "plane of polarization."

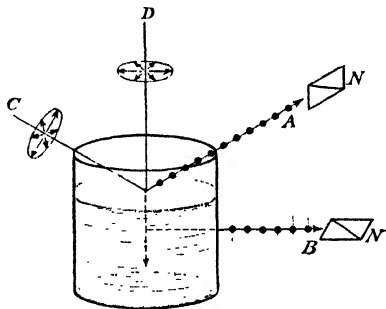


Fig. 126.

518. Interference of polarized light. Like ordinary light, polarized light may be made to interfere provided, first, that the interfering beams originate in the same plane polarized beam from a common source; second, that the interfering beams differ in phase by half a wave length; and third, that they are *polarized in the same plane*. To produce these conditions, light from a monochromatic source may be

polarized by passing it through a Nicol's prism. This light may then be split up into two plane polarized beams by any doubly refracting medium. These beams, on emergence, necessarily differ in phase because of their different velocities, and their vibrations are in planes at right angles to each other. Components of each of these, vibrating in a single plane, are finally obtained by using a second Nicol's prism, and are therefore in a condition to interfere, provided the phase difference is a suitable one.

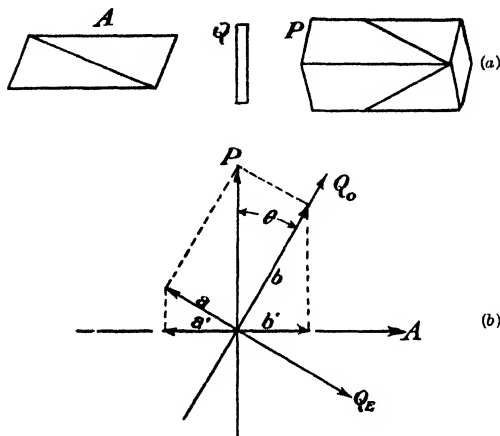


Fig. 127.

If a plate is cut from a quartz crystal so that the optic axis is parallel to the surface of the plate, the maximum retardation of the extraordinary ray is obtained, provided the incident light falls perpendicularly on the plate. In Fig. 127 (a), a plate of quartz Q , cut as specified, is shown placed between two "crossed Nicols," the analyzer

A being turned through 90° with respect to the polarizer P . Their planes of vibration are indicated in Fig. 127 (b) by the vectors A and P . Let the optic axis of the quartz lie in the plane whose trace is Q_E ; then the ordinary ray transmitted by the quartz plate vibrates in the plane represented by Q_O (that is, polarized in the plane of the optic axis), and the extraordinary ray vibrates in the plane defined by Q_E . The beam from the polarizer is decomposed by the quartz into two component vibrations a and b , parallel to Q_E and Q_O , and obtained by projecting P upon these vectors. These components emerge from the quartz and pass into the analyzer, which admits only their components parallel to A . Therefore the vectors a' and b' , obtained by projecting a and b on A , are in a condition to interfere, provided they differ in phase by an odd number of half wave lengths. It will be noted that they are already in opposite phase regardless of path difference, as indicated by the direction of the arrows. This results from the double decomposition, which is equivalent to half a wave length difference of optical path. Therefore, if they are to

interfere, the extraordinary ray must be retarded in its passage through the crystal by an even number of half wave lengths. Thus if the path difference is λ , the vibrations a' and b' differ by $\lambda + \lambda/2 = 3\lambda/2$, which is an odd number of half wave lengths, and interference results.

519. Colors due to thin crystal plates. Beautiful colors are produced when thin sheets of selenite or mica are placed between two Nicol's prisms. In this case, white light is used, and the vector diagram for uniaxial crystals is the same as in Fig. 127 (b), though θ should be 45° for maximum brightness. The relative retardation of the two rays in passing through the crystal is different for different colors, violet being the most retarded, as in ordinary refraction. Suppose a thin plate of selenite is placed between two crossed Nicols with its optic axis at 45° from the plane of the polarized beam. Let the thickness of the crystal be such that the path difference for green light, owing to retardation, is an even number of half wave lengths. Then green will be completely destroyed, because of the extra half wave length explained above, when the beams are recombined by the analyzer. The resulting light is white minus green, and appears of a reddish hue. If red had been destroyed, the resulting color would be greenish. Destruction of yellow leaves a bluish tint, and so, for any color destroyed, the residue of the spectral colors gives a beautiful tint due to their combination. If θ is not exactly 45° , the colors are less brilliant, and if the Q vectors coincide with P and A , no light is transmitted.

In the foregoing arrangement, the field of the interference colors is black. That is, if the plate is not as large as the beam in which it is placed, it has a black background, because the Nicols were crossed. But if they are parallel, as in Fig. 128, the background is white, although the quartz plate may still give interference colors. In this case, however, the conditions for interference are not the same. If

P is projected upon Q_o and Q_x as before, and these components are projected upon A , the resulting vectors point in the same direction, and are equal if $\theta = 45^\circ$. There is now no phase difference due to this resolution, and if the rays are to interfere, the path difference in the crystal must be an odd number of half wave lengths, instead of even, as before, and that color which was previously destroyed with crossed

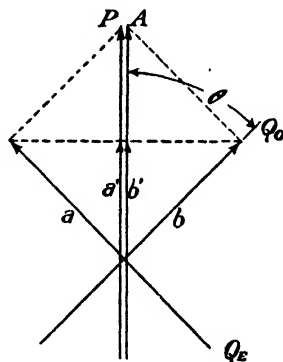


Fig. 128.

Nicols is now most freely transmitted. This results in a hue which is "complementary" to that previously produced by the same crystal plate, because the two colored beams would, if combined, produce white light, what is missing in one being present in the other. They are then said to complement each other, as will be more fully explained in the next chapter.

520. Rotation of the plane of polarization. Let a plate of quartz be cut with the optic axis parallel to the faces of the plate, and further, let it be of such a thickness that when a beam of polarized monochromatic light is sent through it, the extraordinary rays are retarded half a wave length more than the ordinary. Then such a plate has the power of rotating the plane of polarization of the incident light, and it is known as a "half-wave quartz." The angle through which

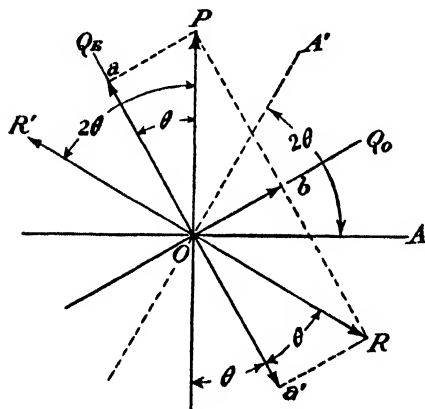


Fig. 129.

the plane of polarization is rotated depends upon the direction of the optic axis of the quartz with reference to the plane of vibration of light from the polarizer, as will be seen from the following explanation: In Fig. 129, if the optic axis of the quartz, defined by Q_E , makes an angle θ with P , the beam from the polarizer is resolved into two components, a and b . But on emergence, a has been reversed in phase with reference to b , because of the half wave

length difference due to the plate's thickness, and it is now represented by a' . Then b and a' recombine to form a resultant R whose plane makes the angle 2θ with respect to P . This is proved as follows: The triangles aPO and $a'RO$ are equal by construction. Then the $\angle POa = \theta = \angle ROa'$. But $\angle R'Oa = \angle ROa'$, since they are vertical angles. Therefore $\angle R'Oa = \theta$, and $\angle R'OP = 2\theta$.

521. Rotatory polarization. Certain media possess the property of rotating the plane of polarization by an amount dependent upon the thickness of the medium. This is quite different from the rotation produced by the half-wave quartz, where the amount of rotation is determined by the position of the optic axis, and where the thickness cannot be varied.

The angle α through which the plane of polarization is rotated depends only upon the wave length with a given thickness of the substance. If the source is monochromatic, an analyzer set with its plane of transmitted vibrations perpendicular to those from the polarizer will extinguish the light before it has entered the substance, but it must be set at a new angle, $90^\circ + \alpha$, to extinguish the light after passing through it. If white light is used, no definite plane exists, for the different wave lengths are rotated through different angles. The amount of rotation varies nearly as the inverse square of the wave length, and when the analyzer is set to extinguish the red light in the emergent beam, the remaining colors are more or less freely transmitted, giving a complementary hue. Similarly, advancing the analyzer to extinguish yellow or any other color leaves a tinted residuum. Therefore the *rotatory power* of a medium has no meaning and cannot be measured, unless light of a specified wave length is used. This is usually light of the *D* lines of sodium.

Quartz plates cut normal to the optic axis, and other double-refracting crystals, have the property of rotating the plane of polarization. Some cause a rotation which is clockwise when viewed toward the source of light. These are called **dextrorotatory**. Some cause a counterclockwise rotation, and are called **laevorotatory**. A few substances such as quartz are found to have both kinds of rotation in different crystals. Some liquids such as turpentine, solutions of Rochelle salts, quinine sulphate, nicotine, camphor, and sugar are also "optically active," as it is called. The most important of these are the various sugars, such as dextrose (cane sugar), which is dextrorotatory, and laevulose (fruit sugar), which is laevorotatory.

522. The Laurent saccharimeter. The measurement of the optical rotation produced by sugar in solution is very important in a number of ways. Sugar appears in the urine of those afflicted with diabetes, and the amount present may be accurately determined by the degree of rotation produced. Import duties on liquids containing sugar may be based on the degree of concentration, and this is most easily and accurately measured in the same way. And in general, researches in chemistry and physics often depend upon an accurate measurement of the degree to which a medium like sugar is optically active.

It would seem as if all that is necessary to determine the angle of rotation would be two Nicol's prisms as polarizer and analyzer, a monochromatic source of light, and a tube to contain the solution. But this arrangement is not at all sensitive, as it is impossible to determine with precision the exact angle at which complete extinction occurs.

Several polarimeters (called **saccharimeters** when graduated to read sugar concentrations directly) have been constructed to give accurate measurements of the angle α . The most important of these were designed by Laurent and by Soleil. The latter instrument is the more delicate, but for most purposes the less complicated and less expensive Laurent type gives sufficiently accurate results.

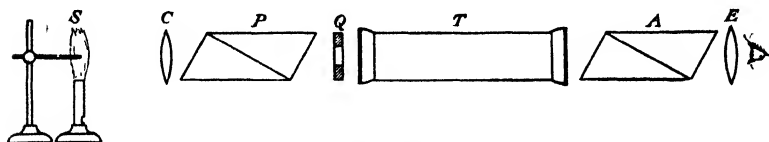


Fig. 130.

In addition to the two Nicol's prisms, the Laurent saccharimeter has a collimating lens to produce a parallel pencil of light, an eyepiece for close observation, and a half-wave quartz plate which partly intercepts the light before it enters the column of liquid to be examined. The optical train is shown in Fig. 130, where S is the monochromatic source, usually a sodium burner, C is the collimating lens, P the polarizer, and Q is a half-wave quartz plate with a hole in its center, or half of a disc of the half-wave quartz. The tube T is fitted with plane-glass covers at its ends and contains the solution

whose rotatory power is to be measured by turning the analyzer A as observed through the ocular E .

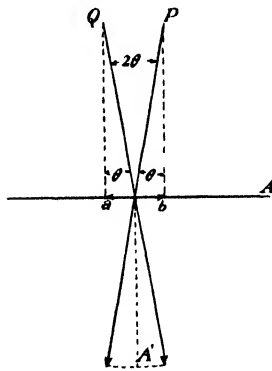


Fig. 131.

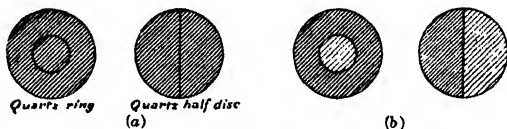
If the optic axis of the quartz ring (or half disc) is set at an angle θ with the vector P , the beam which goes through the quartz is rotated through 2θ , as indicated by the vector Q in Fig. 131. If T were inactive and if the analyzer were set in the position indicated by the vector A , so as to make equal angles with P and Q , the light which has been rotated and that which was not would appear equally intense, as indicated by the equal components a and b , and the quartz

ring would appear just as bright (or dark) as the hole through which the originally polarized light passes without alteration. If the angle θ is small, the two transmitted beams would be almost extinguished and equally dark, but a very slight shift of A makes one brighter and the other darker. The eye is very sensitive to such gradations when

placed side by side, and it can judge of equality with sufficient precision to read the setting of A , on a scale fitted with a vernier, to a tenth of a degree, or even better.

Now let the tube T be filled with an optically active liquid which rotates both P and Q by equal amounts, depending upon the length of the tube and the nature and concentration of the liquid. This angle is found by making a setting of A , first with distilled water in the tube, and then with the solution in question. The difference of these readings gives the value of the angle of rotation. Care should be taken to avoid setting the analyzer so as to bisect 2θ in the position A' . Then the two components of P and Q , projected on A' , are also equal and very bright. But this setting is highly insensitive because these projections are proportional to the cosine of θ , and the cosines of small angles are very insensitive to small changes of the angle.

When the analyzer is set correctly, with distilled water in the tube, the field appears as in Fig. 132 (a). Then when the solution is introduced, the balance is disturbed, and the field appears as in (b). But by rotating the analyzer, the equality of the components a and b is restored, the field once more appears as in (a), and the angle α may be read by the vernier.



523. The Faraday and Kerr effects. In 1845, Michael Faraday, the eminent British physicist and pioneer in the realm of electromagnetism, discovered that polarized light, in passing through isotropic media such as glass, was affected by a magnetic field which rotates the plane of polarization. This effect is most pronounced in media having a high index of refraction, so that flint glass ($n_D = 1.65$) or carbon bisulphide ($n_D = 1.63$) gives a much better result than ordinary glass, or water ($n_D = 1.33$).

In demonstrating the Faraday effect, the substance to be examined is placed between the poles of a powerful electromagnet pierced with holes so that the light may pass through them parallel to the lines of force. In Fig. 133 are shown the essential parts of the apparatus for examining the magneto-optical rotation produced by the magnetic field acting on the medium M . This is placed between the poles N and S as shown, and a Laurent optical system is provided, having a polarizer P , half quartz Q , and analyzer A . The analyzer is first set so as to make the field of uniform brightness with the sample M in

place, but with no magnetic field. Then the switch K is closed and the battery B excites the magnet. This unbalances the optical setting as seen from E , and the analyzer must be turned through some angle θ to restore the two parts of the field of vision to an equal shade. The angle through which it is turned measures the amount of the rotation, and its direction is found to be clockwise when viewed from the source, provided the magnetic field is directed the same way (N to S) as indicated in the diagram. This is like the rule of the right-handed screw, which moves away from the observer when turned clockwise.

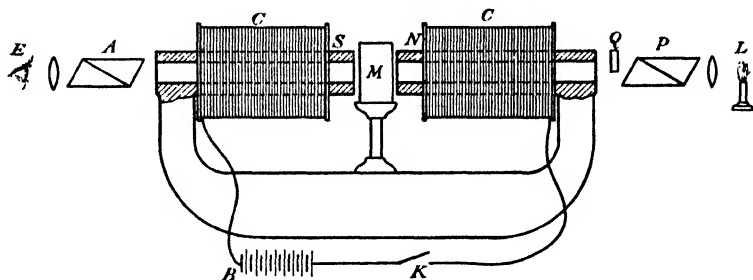


Fig. 133.

If now the field is reversed, the sense of the rotation of the plane of polarization reverses also. From this it is evident that such a beam, if reflected back through M against the magnetic field, would experience twice as great a rotation as after a single passage, because with its direction reversed, the rotation would appear counterclockwise from the analyzer, when directed *against* the magnetic field. But this is equivalent to a clockwise rotation for an observer at L . Thus the total rotation is doubled.

When polarized monochromatic light passes through a given substance, the angle of rotation produced by the magnet depends upon the strength of the magnetic field and the length of path through M . But it depends also upon the wave length of the monochromatic source, in the same way as in passing through a sugar solution, or any substance capable of rotating the plane of polarization. That is, the angle varies nearly as the inverse of the square of λ , as stated in Article 521.

Faraday tried to detect the influence of an electrostatic field on the plane of polarization, but failed to do so. This effect was ultimately discovered by Dr. Kerr of Glasgow, who in 1875 announced the result of an experiment which showed that plane polarized light passing

through a dielectric became elliptically polarized under the influence of a strong electrostatic field. This means that the vibrations occur in elliptical orbits, instead of in a straight line as in plane polarized light. Kerr discovered also that plane polarized light, when reflected from the polished surface of a very powerful electromagnet, becomes elliptically polarized.

SUPPLEMENTARY READING

J. Valasek, *Elements of Optics* (Chap. 12), McGraw-Hill, 1928.

J. K. Robertson, *Introduction to Physical Optics* (Chapters 12, 13, 14), D. Van Nostrand, 1935.

R. W. Wood, *Physical Optics* (Chapters 9 and 10), Macmillan, 1934.

PROBLEMS

1. Calculate the polarizing angles of light crown glass, dense flint glass, and rock salt with sodium light, using Brewster's law. *Ans.* $56^{\circ} 36'$ (crown), $58^{\circ} 47'$ (flint), $57^{\circ} 5'$ (rock salt).

*2. The ordinary index of refraction for sodium light in calcite (Iceland spar) at 18° C is 1.6584. The extraordinary index is 1.4864. If a thin lamina of this crystal is cut with the optic axis parallel to the plane of the lamina and placed between crossed Nicols, what is the minimum thickness needed to produce destructive interference of sodium light? *Ans.* 3.4 microns.

CHAPTER 40

Color

524. Colors of nonluminous objects. When an object is not self-luminous, it can be seen only by reflected or transmitted light. A perfectly white surface reflects all the colors of the spectrum with the same relative intensity as it receives them, even if it absorbs a small percentage of each. In this sense an ideal mirror is a white object, but in general the term *white* is applied only to surfaces rough enough to scatter the light in all directions, so that no mirror images are formed, and when illuminated with white light, the surface appears white in any position. Black objects absorb all colors in equal proportion, though a black surface may reflect a small percentage of the light it receives and still appear black. The difference then between black and white is one of degree only, and with more or less complete reflection or absorption they merge into one another imperceptibly through tones of gray.

Perfectly transparent bodies are invisible when surrounded by a medium of the same refractive index. A piece of flint glass disappears when immersed in carbon bisulphide, both having nearly the same index.

Partially transparent bodies, such as colored glass, may transmit one color freely while absorbing all others. Red glass appears red when white or red light shines through it, but is opaque to light which lacks red rays.

The color of opaque bodies is due to the same atomic properties which enable transparent bodies to color a beam of white light. But in this case the light penetrates only a short distance below the surface and then comes out again as a reflected beam, having lost the same colors by absorption that would be lost by transmission. Red glass reflects red light just as it transmits it, while most opaque bodies, cut extremely thin, transmit the same color they reflect.

In the process of reflection, some of the reflected light has not penetrated below the surface, and a certain percentage of incident white light remains white and then dilutes any reflected color. This dilution is more pronounced if the area of the colored surface is in-

creased by scratches or by pulverizing the material. Crystals of copper sulphate when pulverized become almost white, and red lead becomes pale if it is ground into very fine particles. This is an important consideration in mixing pigments, which are minute particles of some colored material like red lead, mixed with an oil or varnish that acts as a "vehicle." In order to prevent the reflection of white light from the surface of these particles with consequent dilution of color, the vehicle should have approximately the same index of refraction as the particles.

525. Mixing colors. The colors obtained by mixing pigments are due to that portion of the spectrum which both pigments reflect while absorbing all the rest between them. Thus cadmium yellow absorbs blue and violet; French ultramarine absorbs red, orange, and yellow. The mixture then appears bright green because the pigments agree in reflecting this color with remarkable purity, but between them they absorb all the rest of the spectrum. Mixing pigments obviously involves *subtraction* of colors, and is quite unlike the *addition* of colors, to be described later.

If an object is to appear colored, either by reflected or transmitted light, the light it receives must contain that color. If not, the object appears black or dark gray. A green object appears black in a photographic dark room illuminated by a "ruby lamp," but a piece of white paper looks red. Letters written on white paper with red ink are almost invisible under the ruby light, because the ink reflects red light as freely as the white paper. Thus it is evident that the colors of objects depend upon the color of the illumination, and when seen under a mercury or neon lamp, may appear totally different from their appearance by daylight.

526. Surface color. Some substances appear differently colored according to whether they are seen by transmitted or reflected light. Such bodies do not owe their color to selective absorption, but to selective reflection. In this case the surface reflects light of a particular wave length very powerfully, and the rest is transmitted if the body is transparent, or absorbed if it is opaque. Fuchsin, an aniline dye, is one of these. Thin films of gold are yellow by reflected but green by transmitted light, while white light shining through a thin film of silver is colored blue. The spectra of substances having surface color show strong absorption bands due to resonance of their atoms to certain well-defined wave lengths.

527. Color combinations. Let a spectrum of white light be formed as was shown in Fig. 77, but with the slit at the principal focus of the

lens. Then place a diaphragm D , having a rectangular opening, in the path of the refracted beam, as shown in Fig. 134. A lens L properly situated would form real colored images of the slit appearing as a continuous spectrum on a screen at P . But if the screen is moved out to S , where L forms a real image of the rectangular opening, this

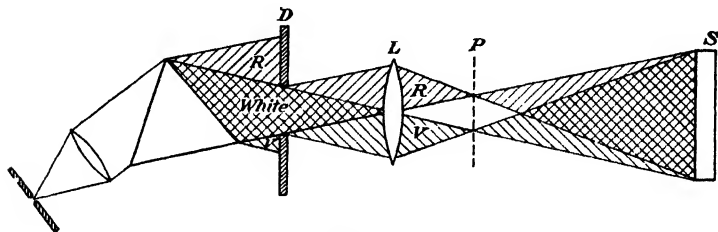


Fig. 134.

image must be white, because all colors pass through the opening in D and are all spread over the entire image at S . Thus the lens recombines the colors split up by the prism.

With the arrangement just described, we may perform a number of interesting experiments by interposing an opaque screen in the plane P , so as to cut off different portions of the spectrum, which are recombined at S . Thus if the screen cuts off red, orange, and yellow, as indicated in Fig. 135, the rectangular image at S will be colored by the

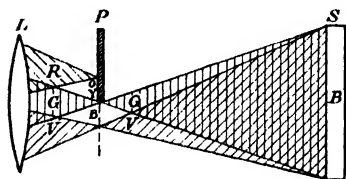


Fig. 135.

sum of green, blue, and violet only, and a bluish hue is observed. On the other hand, cutting off the green, blue, and violet by shielding the lower half of the spectrum at P results in an orange-red hue at S . If we cut out the central colors, yellow and green, the result at S is

purple-violet, while cutting out the two ends and transmitting the central portion produces greenish yellow at S .

The contrasting pairs of colors just described are called **complementary colors**, because their sum is obviously white. Thus the blue color due to green, blue, and violet, is complementary to orange-red, formed by the remaining portion of the spectrum. Any two colors whose sum appears white are complementary colors, and, as we shall see, a great variety of these pairs is possible. The most important are green and purple, red and blue-green, and orange and blue. These pairs furnish violent contrasts and are used in decoration wherever

this effect is desired. On the other hand, adjacent colors of the spectrum, like green and yellow, red and orange, or green and blue, suggest each other and do not startle the eye when placed side by side.

528. Addition of colors. In the experiment just described, the hues were obtained by the addition of the colors contained in a broad spectral band, but the addition of approximately monochromatic colors produces other colors in the same manner. Thus if an opaque screen at *P* (Fig. 135) has fairly narrow slits which allow only the red and the green of the spectrum to pass, the rectangular image at *S* will appear yellow, although no light of the wave length of spectral yellow falls there. Similarly, slits which admit only spectral green and violet result in greenish blue, while red and violet produce purple or magenta.

Like yellow, green-blue has a place in the spectrum with a definite wave length, which, of course, is lacking when it is produced by adding green and violet. But the purple formed by the sum of the extreme ends of the spectrum is a nonspectral color and has no single characteristic wave length.

If the screen at *P* has three slits admitting narrow bands of red, green, and violet, the result is white, which is indistinguishable from the normal white composed of all the colors.

The results of the preceding experiments may be summarized in the following equations:

$$\text{Green} + \text{Violet} = \text{Green-blue.} \quad (1)$$

$$\text{Violet} + \text{Red} = \text{Purple.} \quad (2)$$

$$\text{Red} + \text{Green} = \text{Yellow.} \quad (3)$$

$$\text{Red} + \text{Green} + \text{Violet} = \text{White.} \quad (4)$$

By subtracting (1) from (4) and transposing green-blue, we obtain the complementary relation

$$\text{Red} + \text{Green-blue} = \text{White.} \quad (5)$$

Similarly, combining (2) with (4), and (3) with (4) gives respectively

$$\text{Green} + \text{Purple} = \text{White.} \quad (6)$$

$$\text{Violet} + \text{Yellow} = \text{White.} \quad (7)$$

529. The perception of color. It is clear from the facts stated above that the sensation of color is quite a different concept from the light of the particular wave length or group of wave lengths that produces it. Sensation is a subjective experience, while a beam of light, whether seen or not, is purely objective. A beam of light having a wave length of 0.56 micron may be called yellow for conven-

ience, but a beam of red (say 0.72 microns) combined with one of green (say 0.52μ) *looks* yellow, although objectively there is no yellow in the combination. Therefore care must be taken to distinguish between color sensation and "color" used for convenience to denote certain portions of the objective visible spectrum. These portions may be roughly defined as lying between the following wave-length limits, measured in microns (10^{-4} cm):

Violet	0.33 to 0.45	Yellow	0.55 to 0.59
Blue	0.45 to 0.49	Orange	0.59 to 0.63
Green	0.49 to 0.55	Red	0.63 to 0.81

These six colors are divided into minor gradations sufficiently distinct to admit of about 130 recognizable shades, but combinations of the six colors produce other color *sensations* numbering tens of thousands, many of which have been more or less accurately classified.

530. The Young-Helmholtz theory of color vision. The best-known theory of color vision was originally proposed by Thomas Young (1773–1829), a British physician and natural philosopher of remarkable ability. He was the first to postulate the transverse vibration of light, and in 1802 discovered the fact that any color can be produced by the combination in suitable amounts of only three so-called "primary colors," and suggested a physiological explanation. But Young himself did not attach much importance to his theory, and it was not until the German physicist von Helmholtz developed it, that the scientific world became interested in Young's ingenious hypothesis.

The primary color sensations first proposed by Young were red, yellow, and blue, having wave length ratios of 8:7:6. But in the same year (1802), acting on a suggestion made by Wollaston, he adopted red, green, and violet, having ratios of 7:6:5. This triad is undoubtedly correct, if the three-color vision theory is the right one. However numerous other triads can be used to produce all known colors, for as Parsons† puts it, "any three lights are suitable, provided that neither one can be matched by a mixture of the other two."

To account for compound color sensations, Young suggested that the eye might be capable of responding to only three fundamental kinds of stimuli, each excited by one of the three primary colors. These sensations, combined in varying proportions, were assumed to give rise to all our color sensations.

† Sir John H. Parsons. "Thomas Young Oration" (1931). *Trans. Opt. Soc.*, Vol. 32, pp. 165–185.

The chief physiological aspects of color vision are as follows: At the back of the retina, away from the pupil, are thousands of minute bodies, packed closely side by side, known as rods and cones. They are perpendicular to and almost in contact with the black choroid membrane, which lies between the sclerotic and the retina. The rods are most plentiful in its outer margins, where color is least appreciated, and it is felt that twilight vision (almost devoid of color) is due to the rods that perceive only light-dark values. The cones, on the other hand, are most plentiful in the yellow spot, where color is most vividly perceived. These cones are supposed to contain the three types of receiving apparatus (nerve ends?) necessary to perceive the three primary colors, but according to the theory, each is not wholly insensitive to the other two colors and may be partially stimulated by them.

To account for the known facts of addition of colors, complementary colors, and so forth, sensitivity curves have been constructed in accordance with the available data.

Probably the best are due to Koenig, and these are shown in Fig. 136, where the ordinates are proportional to sensitivity, and abscissae to wave length. The high peak of the

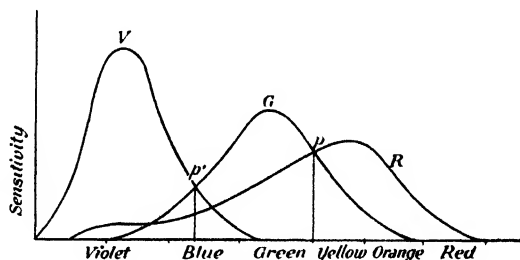


Fig. 136.

violet curve shows the observed fact that the eye is most sensitive to this color. This is doubtless because the energy received on the earth from the sun falls off very rapidly toward the shorter wave lengths, although it passes through a maximum in the greenish yellow.

531. Interpretation of the sensitivity curves. The reason why we see white light when monochromatic red, green, and violet are added, is shown by the spread of the curves over all visible wave lengths. The addition of red and green would appear greenish yellow, because that color alone would stimulate the two sensations equally, as at the point p in Fig. 136; therefore an equal stimulation of each independently at the peaks of the curves produces nearly the same sensation as a joint stimulation by a single wave length at p . In a similar manner, a suitable stimulation of the violet and green sensations gives the impression of blue, corresponding to the point p' .

Complementary colors are also easily explained. Suppose, for instance, violet light of suitable intensity is added to greenish yellow. The greenish yellow stimulates the red and green sensations about equally, and the added violet supplies the third component necessary for the sensation of white, barring twilight white seen by the rods. In a like manner, blue light, which stimulates both the violet and green sensations about equally, is complementary to yellow, because spectral yellow stimulates red strongly and green much less so.

It is not surprising that Koenig's sensitivity curves explain the facts so well, because they are drawn so that they should do so, and we get out of them precisely what we put in. Though there is some physiological evidence that the three independent types of receiving organs really exist, it is better to regard Koenig's curves only as a convenient way of expressing what actually happens.

Whatever may be the ultimate mechanism of color vision, it is certain that the eye has a wonderful power of combining colors to form new sensations, some of which are radically different from their components. This *synthetic* power is characteristic of the eye, just as the power of *analyzing* compound tones is characteristic of the ear. That is why, as Mrs. Ladd-Franklin† has observed, "we can never have in the play of colors intricate aesthetic combinations and involutions corresponding to musical compositions in tones."

532. Hering's color theory. There are several serious difficulties in the way of the Young-Helmholtz theory, such as the two kinds of yellow (spectral yellow, and red + green); achromatic vision, which is unaccounted for by a strictly trichromatic theory; and **afterimages**. These are the well-known colored spots one sees after staring at a bright object for some time. If the object is red, the afterimage, when seen against a white background, is the complementary blue-green color. These difficulties are better met by a theory originated by Ewald Hering (1834-1918), a German physiologist.

But before discussing Hering's theory, we should take note of an important factor in vision known as the **visual purple**. This is a colored substance contained in the ends of the rods next to the choroid and farthest from the pupil. It turns white under exposure to light, but in darkness its color is restored from a layer of pigment cells between the retina and the choroid. Apparently it is the visual purple which gives us our achromatic sensations of light and dark through the agency of the rods. If we have fatigued the eye by a bright light, the visual purple is temporarily exhausted and the retina is very

† Ladd-Franklin, *Colour and Colour Theories*, Harcourt, Brace, 1929.

insensitive. But rest in a dark room restores this material, and normal sensitivity is regained. It takes four or five minutes in perfect darkness to acquire the maximum supply of visual purple and maximum sensitivity to light.

The behavior of the visual purple just described represents the processes of **assimilation** and **dissimilation** which go on continuously in all living organisms, a building-up and a simultaneous breaking-down. Hering used this principle in explaining not only light-dark vision, but also that of colors. In the case of achromatic vision, the sensation of black is associated with the assimilative process (*A*), while white is associated with dissimilation (*D*). Thus with this particular pair of opposites, any light at all, such as gray, is dissimilative, and causes the bleaching of the visual purple.

In addition to black and white, Hering based his theory on two pairs of opposition colors (*gegenfarben*). These are red and green, and yellow and blue, colors which convey no suggestion of each other. This then is a tetrachromatic theory having four primary colors instead of the three proposed by Young.

According to Hering there are a red-green process and a yellow-blue process, in addition to the white-black process already described. The red-green process is of the *D* kind, from the shortest visible wave lengths to pure blue, giving the sensation of red. Then it reverses, and from blue to pure yellow it is of the *A* kind, giving a green sensation. From yellow through red it reverts to the *D* kind and again causes the sensation of red. The yellow-blue process is of the *A* kind, from extreme violet to pure green, giving a sensation of blue. From green to the extreme red it is of the *D* type, and the resulting sensation is yellow. The intensities of these sensations vary with the wave length, and may be represented by the two curves shown in Fig. 137. Here we see that the four primary colors are produced by the action of a single one of the two processes, for in each case the curve representing the other is passing through the neutral line, indicating no sensation when that particular wave length acts upon it.

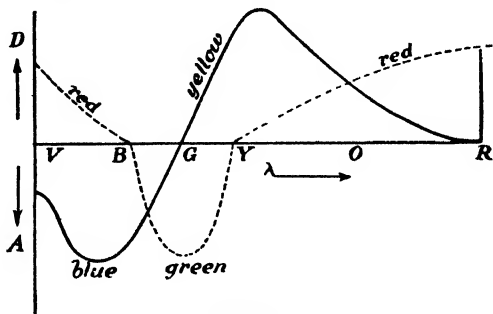


Fig. 137.

The other colors, such as violet, blue-green, and orange, are combination colors due to both processes acting simultaneously.

Hering's theory is more popular with painters, physiologists, and psychologists than it is with physicists, who regard it as somewhat artificial, unsatisfactory in accounting for the observed facts of color blindness, and almost certainly wrong in regarding yellow as a primary color. Their attitude may be summed up in the words of Sir John Parsons, who concludes that it is "unwise in the present state of knowledge to abandon the methods of examination founded upon the three components theory."

533. Classification of colors. Maxwell greatly extended the Young-Helmholtz theory by means of his famous color disc. This is made up of sectors of variable angles and hues. When rotated at a high speed, the colors of the sectors merge into a single hue, because so-called *persistence* of visual impressions enables the eye to combine them. The resulting sensation is the same as that caused by true color addition, described in Article 528. In this way Maxwell succeeded in matching any desired hue by a combination of two or more primary colors, and a formula was obtained expressing the hue in terms of its ingredients. Such a formula is

$$C = x_1S + x_2G + x_3U, \quad (1)$$

where S , G , and U stand for Dominguez's primary colors†, scarlet, green, and ultramarine, and x_1 , x_2 , and x_3 are the **valences**, or color co-ordinates, of the combination color C . These valences are expressed as numbers whose absolute values are meaningless, but their ratios determine the proportion of the three ingredients needed to match a given color. Thus $6S + 2G + 7U$ would define a particular hue.

When two of the valences are zero, C is a primary color. When one valence is zero, C is one of the duplex colors—yellow, turquoise (green-blue), or purple. When $x_1:x_2:x_3 = 1:1:1$, the sensation is arbitrarily taken as white. Theoretically, any valence can be negative, but as a negative color sensation is meaningless, we have to transpose the negative term to the other side of the equation, giving, for instance,

$$C + x_1S = x_2G + x_3U, \quad (2)$$

which is a perfectly rational statement.

† "Investigation on Impure Spectra," by C. Villalobos Dominguez, of the University of Buenos Aires. Dominguez's primary colors are more correctly named than those of earlier writers.

In addition to **hue**, there are two other considerations in color classification. These are **saturation** and **luminosity**. Saturation determines the purity of a color, which is diminished by an admixture of white light. Spectral colors are obviously the most saturated of all colors, provided we include with them the purples, which are just as "pure" as the duplex colors of the spectrum. Colors, however, such as rose, straw, and lavender, are not found in the spectrum and are not saturated. These are obtained respectively by diluting red, yellow, and violet with white; therefore none of the three valences can be zero, and such colors are triplex.

Degrees of luminosity or brightness affect our perception of color in another way. Colors of low luminosity are called dull, as "dull green" or "dull blue," which are decidedly different from the bright colors. In general, these dull colors have no specific name, except that with increasing dullness all alike end in black. But dull yellow and dull orange have such a special character that they have received the distinctive name of *brown*.

SUPPLEMENTARY READING

M. Luckiesh, *Color and its Applications*, D. Van Nostrand, 1915.

J. P. C. Southall, *Mirrors, Prisms and Lenses* (pp. 720-747), Macmillan, 1933.

A. H. Munsell, *A Color Notation*, Munsell Color Co., Baltimore, 1926.

CHAPTER 41

Sources of Light

534. Incandescent solids. When a bar of iron is heated in a furnace, as its temperature rises it emits radiant energy over a range of frequencies having an upper limit that rises with the temperature. As it grows hotter, there is an increasing amount of energy of all frequencies below this rising limit, very much like the behavior of water going over a dam. When the crest of the fall is just above the dam, the water falls almost vertically, but as the height rises, some of the water is carried farther and farther from the base of the dam, while an increasing amount falls within this limit.

At a temperature a little above 400°C (405° according to Emden), the wave length of some of the emitted radiation is sufficiently short to affect the eye, and the iron glows with a dull light called "gray glow," just barely visible in the dark. When made still hotter, it reaches ordinary "red heat" (about 525°), then becomes "cherry red," then orange, and then yellow at about 1000°C , and finally, above 1200°C , it is "white hot." This white heat marks a somewhat indefinite temperature, because it may still have a yellowish hue at a very high temperature. However, as the amount of blue and violet in the spectrum of the glowing iron increases, its light becomes more and more like daylight, and begins to approach the blue whiteness of the electric arc, whose positive crater has a temperature exceeding 3500°C .

Different metals begin to be luminous at different temperatures, according to their emissivity. Gold begins to emit visible radiations at 423°C , and platinum at 408°C , while the temperatures for other metals are not very far from these values.

Like water going over a dam, the energy emitted by a glowing solid is not evenly distributed throughout its spectrum. In the water analogy, there is a certain range within which most of the water falls. So in the case of the heated body more energy is emitted at a certain frequency than at any other. This frequency, which corresponds to the maximum energy, rises with the temperature, just as the upper limit already referred to rises, or, what is the same thing, the wave

length which is most freely emitted (where the spectrum is most intense) becomes progressively shorter.

535. Wien's displacement law. In 1893 it was shown by Wien, a German physicist, that the wave length which characterizes the maximum of energy distribution in the spectrum of a radiating body varies inversely as its absolute temperature. This may be written $\lambda_m = K/T$, where K is a constant and has a value of 2885 for black-body radiation, provided the absolute temperature T is measured in centigrade degrees, and λ_m in microns (10^{-4} cm). This relation has also been derived from purely theoretical considerations, and is exact in the case of an ideal black body. With other bodies, Wien's constant is smaller. In the case of platinum, for instance, it is about 2630.

The significance of Wien's law is shown by the curves in Fig. 138. Each one, representing a stated temperature, shows the distribution of the intensity of radiation over a range of wave lengths at that temperature. It will be seen that the energy emitted at a given wave length increases all through the spectrum as the temperature rises, but that the increasing maxima shift steadily to the left in the direction of shorter wave lengths. However, even at 1200° , where a black body is said to be white hot, the peak is well down in the infrared. It is thus evident that the sun, whose distribution curve observed from sea level has a maximum in the yellow-green, must have a very much higher temperature than red-hot iron. In fact, using Wien's formula with $\lambda = 0.55$ microns for yellow light, we obtain $T = 5245^\circ \text{K}$, or 4972°C .† But the sun is not a "black body," and moreover, the shorter wave lengths are largely absorbed by the earth's atmosphere, so that this figure is certainly too low. Allowing for those effects, 6000°C is considered a reasonable value for the sun's surface temperature.

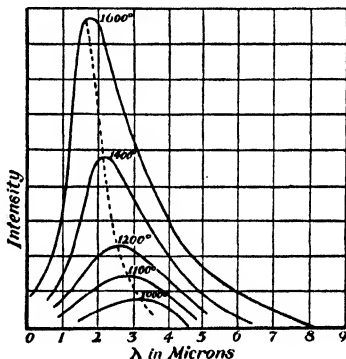


Fig. 138.

The measurement of energy distribution throughout the spectrum of a star is a difficult undertaking, but it has been accomplished by *spectrophotometry*, and in this way surface temperatures of $20,000^\circ \text{K}$ have been observed.

† Compare this value with the one obtained in Article 304 by a different and apparently better method.

536. Planck's radiation formula. Various attempts have been made to find an expression which would make it possible to calculate the energy emitted by a black body at any wave length and temperature, or, in other words, to plot the curves of Fig. 138 from purely theoretical considerations. With this aim in view, both Wien and Rayleigh derived formulae that fitted the experimental curves for some portions of the spectrum, but were not at all satisfactory in other portions. Finally, Max Planck, a celebrated German physicist, derived a formula which is a surprisingly accurate statement of the relations between the energy, the wave length, and the absolute temperature. In the equation given below, I is the intensity of radiation whose wave length is λ . Then

$$I = \frac{hc^3}{\lambda^5 (e^{hc/k\lambda T} - 1)},$$

where c is the velocity of light, k is Boltzmann's constant ($= R/N$) defined in Article 204, and h , known as Planck's constant, equals 6.624×10^{-27} erg-seconds.

537. The quantum. Planck's equation is of great theoretical importance, because it embodies the idea of the **quantum**, which he found necessary in order to account for the observed facts of radiation, and so obtain a formula which fitted them. According to the quantum theory, radiant energy is not a uniform flow, but takes place in minute bundles, called **quanta**, just as matter is not continuous, but is made up of separate atoms. The energy of a quantum is equal to the product of Planck's constant and the frequency ν , or $W = h\nu$. Since ν is not a constant, W can have a great variety of values. Thus there are large quanta and small quanta, just as there are large and small atoms. A quantum appears to remain intact without diminution in magnitude as it traverses space, even from the most distant stars, although the number of quanta crossing a unit area per second must decrease in accordance with the inverse square law. These particles of energy are also called **photons**, and, traveling with the velocity of light, they possess both mass and momentum equal to $h\nu/c^2$ and $h\nu/c$, respectively. The pressure exerted by radiant energy falling on a surface can be calculated from the momentum of the photons. It is equal to $nh\nu/c$, where n is the number of photons striking unit area per second.

The quantum, although it represents a definite frequency ν , does not fit with the older wave theory of light. The wave theory is necessary to account for such phenomena as diffraction and polarization,

where we seem obliged to reason in terms of a continuous flow of energy in the form of transverse waves. The quantum theory, on the other hand, is necessary in explaining some of the observed facts of radiation and absorption of energy by atoms and electrons. These two ideas seem mutually contradictory, but since 1926 the efforts which physicists have made to reconcile them have had a profound effect on the development of theoretical physics. About that time de Broglie and Schrödinger evolved a new theory known as **wave mechanics**, which provides a description of both the wave and particle aspects of light. This is discussed more fully in Article 797.

538. Incandescent gases. Not only ordinary gases, but solids in the gaseous state may be made luminous in a variety of ways, some of which have already been mentioned. Metallic salts, for instance, become volatilized in the Bunsen burner, and emit light in certain definite wave lengths known as their **flame spectra**. But the electric discharge, either as an arc, an ordinary spark, or through an exhausted tube, is the way most commonly used. A continuous discharge, involving considerable current and rather low voltage, between two carbon or metal rods, is called an **arc**. An **arc spectrum** is due mainly to the material of the electrodes, but there may be added the spectra of gases or volatilized metallic salts placed in the path of the discharge. The true spark is a disruptive discharge between terminals, and is characterized by small current and high voltage. This kind of spark spectrum is that of the metal of the terminals, but when the discharge takes place in an exhausted tube containing a rarefied gas, the **spark spectrum** is that of the gas. Both types of spark spectra are similar in character, though they are produced under different conditions.

Flame, arc, and spark spectra of the same substance may differ greatly. Sodium, for instance, yields only the familiar greenish-yellow *D* lines when volatilized in a flame of moderate temperature, but its spark spectrum has many lines. The arc and spark spectra of ordinary metals also differ profoundly. The spark spectrum is produced by the ionized atom, and its character further depends upon whether it is singly, doubly, or still more highly ionized by the removal of one or more outer-ring electrons. The arc spectrum, on the other hand, is produced by the neutral atom.

539. Spectral series. It has been known for a long time that the lines in the spectra of gases show a certain orderliness of arrangement which could not be due to chance. But the frequencies that these lines represent are not related by any law as simple as that which de-

termines the harmonics of a musical tone, and it was some time before this law could be formulated. In 1885 Balmer discovered the relation that governs the arrangement of the visible spectrum of hydrogen. This is now usually expressed as an equation which gives reciprocal wave lengths of the various lines. Reciprocal wave length, or "wave number," really means the number of waves per centimeter of path of the light, and is more convenient than either frequency or wave length for certain purposes. Balmer's equation is

$$\frac{1}{\lambda} = R \left(\frac{1}{2^2} - \frac{1}{m^2} \right),$$

where R , known as the **Rydberg constant**, equals 109,678 in a vacuum, and m is any integer above 2. Thus, setting $m = 3$, we obtain the wave number of the first line in the series; $m = 4$ gives the second, and so on. These lines are rather far apart at first, but crowd closer and closer together as m increases. The limit of the series is found by making m infinite, when $1/\lambda = 109,678/4$, and $\lambda = 3646 \text{ \AA}$ in vacuo. This is near the extreme violet end of the visible spectrum, although the first, or α , line ($m = 3$) is in the orange red. The agreement between the observed and calculated wave lengths of the series is extraordinary. For instance, the observed wave length of the δ line ($m = 6$) is 4,101.26, and the calculated value is 4,101.30, using $R = 109,721$, which is the value of the Rydberg constant in air.

This series was the only one known in the spectrum of hydrogen for some time, until Theodore Lyman, of Harvard University, discovered another series in the ultraviolet, now called the Lyman series. Its equation is

$$\frac{1}{\lambda} = R \left(\frac{1}{1^2} - \frac{1}{m^2} \right),$$

where $m = 2, 3, 4$, and so forth. Several other series exist in the infrared, the most important having been named after Paschen, who discovered it. It is given by the formula

$$\frac{1}{\lambda} = R \left(\frac{1}{3^2} - \frac{1}{m^2} \right),$$

where $m = 4, 5, 6$, and so forth. Other elements have similar series but they are not so simply described, although the Rydberg constant, which differs slightly among the elements, enters into all of them. Its calculation for hydrogen by Niels Bohr, an eminent Danish physicist, using his famous hypothetical atom, is a remarkable justification of the electronic theory of radiation, to be explained later.

540. Infrared, Hertzian, and "radio" waves. Anything having a temperature higher than the absolute zero is sending out radiant energy, although, as was explained in discussing this question in Article 295, it receives more than it sends out if it is cooler than its surroundings. Those radiations classed as thermal lie in the infrared portion of the spectrum, and are due to the agitation of atoms or molecules as a whole, and not of their component electrons. They are usually examined by means of the bolometer and radiomicrometer. These instruments are very sensitive to small temperature changes, and record them electrically, as will be explained later. Specially prepared photographic plates may also be made sensitive to very long waves, and are used in the infrared region. In these ways, E. F. Nichols and Tear, American physicists, have explored the infrared spectrum of the mercury arc as far as 0.412 mm.

Infrared rays are much used in heat treatment of certain maladies. Those having a wave length of about 11,000 Å are highly penetrating both in water and flesh, and supply internal heat far more effectively than is possible by the slow process of conduction from a hot object applied to the surface of the skin.

Waves longer than those just mentioned are produced by electromagnetic oscillations such as are used in radiotransmission, and there is no theoretical limit to the length that may be produced in this way, although in practice, waves over twenty miles long are not used, and that length only in transoceanic radiotelegraphy. The shortest waves produced by Hertz, who first investigated them, were several meters long, the average being about five meters. These he found could be reflected, refracted, and diffracted like ordinary light, and he thus completely demonstrated the electromagnetic character of light, which had previously been affirmed by Maxwell as a result of purely theoretical reasoning.

Subsequent experimenters, such as Righi and Lebedew, gradually obtained shorter and shorter waves by improved oscillators; in 1897 Sir Jagadis Bose, of the University of Calcutta, obtained waves only 6 mm long, and showed that in Iceland spar they suffered double refraction like ordinary light. These values remained the minimum that could be produced by electric oscillators, until Nichols and Tear closed the gap between thermal and electric waves by producing electric oscillations whose wave length was only 0.22 mm. This is actually shorter than the longest heat waves they had observed.

541. Ultraviolet light and beyond. Heated solids, as we have seen, do not emit much energy in the shorter wave lengths of visible light,

because the necessary temperature cannot be obtained in the laboratory, and if it could, they would no longer be solids. The electric arc, however, and the sun, which is still hotter, do emit large quantities of ultraviolet light a little beyond the visible violet. These are readily observed by photography as far as 2000 \AA (0.0002 mm). But the shorter waves emitted by the sun are strongly absorbed by our atmosphere. Waves much under 2800 \AA do not reach the earth at sea level, and window glass stops waves shorter than 3000 \AA .

The effect of sunlight on the skin, known as erythema (sunburn), is a maximum at about 2950 \AA . As this wave length is stopped by glass, windows of fused quartz transparent to the shorter waves are sometimes used. Sunburn, as Luckiesh† has shown, has no value in itself, but a moderate irradiation of the skin by the near ultraviolet is beneficial.

In the study of these short waves, quartz prisms and lenses must be used, and for still shorter ones it is necessary to work in a vacuum, because air becomes practically opaque in this region of the spectrum. Below 1850 \AA , quartz is no longer available, but fluorite prisms and lenses make it possible to observe waves down to 1200 \AA . Beyond this, the reflection grating must be used to produce the spectrum. The sources of these short waves are lines in the spark spectra of various substances, and Millikan has photographed such lines having a wave length as short as 136.6 \AA .

Beyond the region of what may properly be called ultraviolet light, because it is produced in a manner similar to visible light, we enter the region of X-rays. These are produced by the bombardment of solid metal targets with rapidly moving electrons. Ordinary X-rays, such as were first known, have a wave length of the order of 0.1 \AA , but their actual range today is from 500 \AA to 0.06 \AA . The former are very "soft," and are so readily absorbed that they cannot penetrate the glass of the tube in which they are produced, but must be measured inside it. The very hard rays, on the other hand, have high penetrating power, and can pass through a considerable thickness of metals, including even lead, which completely stops ordinary X-rays.

The **gamma rays** from radioactive substances are similar to X-rays, but are still more penetrating. Their measured wave length extends from 0.5 \AA to 0.006 \AA , which is far beyond the "hardest" X-rays yet produced.

Finally, we should at least mention the famous **cosmic rays**, although their exact character is still in doubt, as we shall see in Arti-

† M. Luckiesh, *Artificial Sunlight*, Van Nostrand, 1930.

cle 781. At any rate, some sort of "rays" come to us from outer space, and their amazing power of penetration indicates either very high frequency, or, if they are particles, enormous speed.

542. Comparison of wave lengths from various sources. The following table gives a comprehensive survey of the various ranges of waves emitted by the sources we have been discussing. It is

Name	Source	General Character	Wave Length
Cosmic.....	Outer space	Can penetrate 70 meters of water or a meter of lead	(?)
Gamma.....	Radium, etc.	Can penetrate a foot of lead	0.006 Å to 0.5 Å
X-rays.....	Electron bombardment	Range from high penetration to "soft" absorbable rays	0.06 Å to 500 Å
Ultraviolet....	Chiefly spark spectra of gases and arc spectra of solids	Highly absorbable even by transparent media	136.6 Å to 3900 Å
Visible light...	Incandescent solids and liquids above 400° C, also gases electrically excited	Affect the retina of the eye. High penetration of transparent media	3900 Å to 7600 Å
Infrared.....	Spectra of gases and "rest strahlen"	Radiant heat. Penetration of many opaque media	7600 Å to 0.412 mm
Hertzian.....	Electric oscillations	More penetrating than "heat" waves	0.22 mm to a few meters
Radio.....	Electric oscillations	Great carrying power and penetration	A few meters to 30 km

evident from these figures that there is now no longer any unexplored region of wave lengths between the longest waves used in radiotelegraphy and the shortest gamma rays emitted by radium. In fact, it will be seen that several of these ranges overlap, so that we may, for instance, obtain waves whose length is 250 Å, either in an X-ray tube or from the spark spectrum of one of the elements. Also, a wave whose length is 0.3 mm can be produced either by the methods of spectroscopy or by electrical oscillations.

Another conclusion we may draw from the table is that both very long and very short waves are extremely penetrating, while those in between approach a maximum of absorbability.

543. Fluorescence. Many substances, though not, strictly speaking, independent sources of light, emit luminous energy on their own account under the influence of a beam of light falling upon them. Those which emit only during the action of the stimulating light are said to **fluoresce**, while those which continue to emit some time afterwards are said to **phosphoresce**. In fluorescence the atoms of the substance absorb a portion of the spectrum of the incident light, and re-emit it in another portion of the spectrum. This transformation is nearly always in the direction of lower frequency and longer wave lengths, a fact known as **Stokes' law**.

Stokes' law may be demonstrated by illuminating with violet light pieces of uranium glass and paper placed side by side in a dark chamber. If these objects are then viewed through amber glass, the uranium glass, which is fluorescent, seems to glow on its own account, while the paper is hardly visible. The reason is that amber glass is almost opaque to the violet light reflected from the paper, but readily transmits the lower-frequency green fluorescence excited in the uranium glass by the higher-frequency violet light. If, however, both objects are illuminated by red light, the paper appears brighter because it is a better reflector, and there is now no fluorescence. Thus fluorescence is usually a sort of "step-down" transformation of frequencies. In accordance with the quantum theory, the energy of the incident quantum must exceed that of the fluorescent light, so that $h\nu_1 > h\nu_2$. Therefore the fluorescent, or secondary frequency, ν_2 , is lower than the primary frequency ν_1 . In order to see this transformed light we must observe it as a reflected or scattered, but not a transmitted beam. This is because the incident light is always much the stronger and masks the much feebler fluorescence, while it is transmitted in its original color slightly modified by selective absorption.

A common example of fluorescence is petroleum oil, such as the heavy engine oils used in automobiles. These appear greenish brown by transmitted light, but the reflected or scattered light has a bluish fluorescence, due to the absorption of still shorter waves in the violet. This violet, taken from the incident white light, gives the transmitted beam a green-yellow hue of low luminosity, in accordance with the principle of complementary colors.

A solution of fluorescein appears yellowish by transmitted light, but fluoresces with a beautiful pale-green hue. Eosin, which derives its name from the Greek word for *dawn*, gives a gorgeous rosy fluorescence. Quinine sulphate, under the action of ultraviolet light,

fluoresces in the dark with a pale-blue color, but shows no color by transmitted light, because the absorbed ultraviolet cannot affect the color of the visible light. Esculin, easily obtained from horse-chestnut bark by boiling it in a small quantity of water, also fluoresces with a pale-blue light. A solution of chlorophyl (the coloring matter of plants) dissolved in ether fluoresces blood red, but appears green by transmitted light. Uranium glass is an example of a strongly fluorescent solid. Other fluorescent substances are the mineral willemite, fluor spar, certain fossils, and calcite (Iceland spar).

Light, after passing through a sufficiently thick layer of fluorescent material, loses its power to excite fluorescence in the same substance placed in the transmitted beam. This is because the wave lengths needed to excite those particular atoms have all been absorbed.

All bodies which fluoresce within the visible portion of the spectrum are excited by ultraviolet light, because the energy of its quanta is greater than that associated with any part of the visible range, and the step-down process of transformation is therefore always possible. This leads to a valuable method for "seeing" otherwise invisible radiations. Thus we may produce the spectrum of an arc light with a quartz prism and lenses and project it upon a sheet of paper previously coated with a solution of quinine sulphate, moistened with dilute sulphuric acid before using. On this screen the visible spectrum is doubled in length, far beyond the usual limits of the violet end, and the spectral lines of the arc in this ultraviolet region are clearly visible.

Professor Wood has invented a kind of glass which is almost opaque to ordinary light and to most of the ultraviolet, but transparent to the near ultraviolet as produced by the mercury arc. This invisible beam may be detected by placing a fluorescent substance in its path. A piece of uranium glass seems to glow by its own light, invisible streaks of vaseline on paper become luminous, and the teeth and pupils of the eye of an observer are also fluorescent, while white porcelain or a silver coin appears coal black.

Röntgen (X) rays may be made visible in a similar manner by the fluorescence of barium platinocyanide, which gives out a greenish-yellow light under their influence. In practice, the X-rays, after passing through the part of the body which is under examination, fall upon a screen coated with this material, and a shadow picture is formed, visible in the dark.

544. Phosphorescence. When the fluorescent light lasts after the cause has been removed, the medium, as has been stated, is said to be **phosphorescent**, because moist phosphorus glows in the dark.

But very many substances possess this property. A familiar example is Balmain's luminous paint, a sulphide of calcium which glows for hours after having been exposed to a strong light. Other substances possessing this property are the sulphides of barium and strontium, and some of the salts of aluminum, uranium, and platinum, as well as all *solids* which fluoresce. Fluorescent liquids and gases, however, are not phosphorescent.

The explanation of phosphorescence is that some sort of chemical change is produced by the original illumination. This change is unstable and therefore of a temporary nature, so that when left in the dark, the exposed surface gradually reverts to its more stable condition, emitting light as it does so. The rate of this recovery and consequent intensity of the light emitted vary with the temperature as well as with the nature of the substance. Dewar found that a piece of ammonium platinocyanide, if cooled to the temperature of liquid hydrogen and exposed to a strong light, exhibited no phosphorescence when removed to a dark room. But later, as it grew warmer, it suddenly developed a brilliant greenish phosphorescence.

The rate at which the stored-up chemical energy is liberated during phosphorescence may be accelerated by warming the body, as we have just seen, and invisible infrared illumination produces the same result. Wood found that Balmain's paint, which had been kept in the dark for 24 hours after exposure, could have its lost luminosity restored for a short time by the action of the infrared, which produced a greenish light instead of the characteristic blue of the paint. This effect is to be distinguished from fluorescence, because it is not a transformation of the incident light into a lower frequency, but is only a release of the small residual chemical energy under the action of the invisible beam.

Many organic materials are phosphorescent, and certain fungi and decaying wood are phosphorescent apparently spontaneously, owing to slow oxidation which results in the emission of a faint light. A strongly phosphorescent fungus is *Clitocybe illudens* of the family *Agaricaceae*. McIlvaine† says that "the print of a newspaper could be read when held close to the mass [of fungi]." Another illustration of a slow chemical change producing light is the luminous streak made by a phosphorus match when rubbed on a moist surface. Phosphorescence is also excited in many crystalline bodies by the action of cathode rays, and in some of them persists for quite a long time after

† Charles McIlvaine, *One Thousand American Fungi*, Bobbs-Merrill Co. 1912.

the bombardment has ceased. The light produced by these crystals is beautifully colored, some glowing with an emerald-green light, others sapphire blue, and some emit a superb vermilion.

545. Luminescence. This term refers to a source of light not primarily associated with heat. Fluorescence and phosphorescence are therefore cases of the broader term **luminescence**. When luminescence is due to a primary beam of light, as described in the preceding paragraphs, it is called **photoluminescence**. When it is due to electronic bombardment, it is known as **electroluminescence**. When it is excited by infrared rays, as in Dewar's experiment, it may be called **thermoluminescence**.

546. The Raman effect. As we have seen, fluorescence occurs only in a comparatively small number of substances, but in 1928 Sir C. V. Raman, of the University of Calcutta, discovered that all transparent bodies have a somewhat similar power. The substance to be examined is illuminated by a source, such as the mercury arc, giving light of one or more definite wave lengths, and the light scattered at right angles to the original beam is examined with a spectrometer equipped to photograph the spectrum. After prolonged exposure it is found that in addition to the original line or lines due to the source, other much fainter ones are present. Their positions with respect to the line that caused them are perfectly regular and in accord with the quantum theory of atomic absorption and radiation. The strongest of these secondary lines are of longer wave length than the primary line, but some fainter lines are found having wave lengths shorter than their primary. The lines of longer wave length are explained on the basis of the quantum theory of atomic absorption. According to this theory an electron within the atom may be raised from a lower to a higher energy level during the impact of a photon, while the energy quantum of the colliding photon is correspondingly reduced, which results in a lowering of its frequency. The secondary lines of increased frequency are accounted for by supposing that the incident photon can not only give up its quantum of energy to the atom, but under certain circumstances can receive energy. This energy, added to its own, increases the value of the quantum $h\nu$, and therefore the frequency. The increase of frequency when the photon receives energy has the same value as the decrease when the photon loses energy to the atom. It is indicated by the symmetrical position of the secondary lines on either side of the primary, so that we may regard the transaction due to the collision as a reversible one. The atom can either rob the photon of a definite amount of energy during the col-

lision, or impart to it the same amount. In one case the frequency is lowered; in the other it is raised.

547. The mechanical equivalent of light. This is defined as the power, measured in watts, which corresponds to a lumen of luminous flux emitted by a source of white light. The number of lumens is 4π times the candle power, and this is measured by a photometer. The corresponding mechanical power is measured by some such device as a thermopile, after the infrared and ultraviolet portions of the spectrum have been filtered out, so that only the visible portion falls upon the thermopile. This is necessarily a somewhat uncertain measurement, because "white" light is of variable composition, depending upon the nature of the source. However, an average value for the mechanical equivalent of light is considered to be about 0.01 watt per lumen.

If we define the mechanical equivalent of light in terms of a monochromatic source, and choose the brightest portion of the spectrum, that is, the yellow-green, then there is no uncertainty as to its composition, and more accurate results are possible, as well as much higher efficiency. In this case the mechanical equivalent has been found to be only 0.00161 watt per lumen. As the eye is capable of perceiving light whose intensity is only 10^{-13} lumen,[†] it is evident that the flow of energy required to excite the sensation of light is extremely small, being of the order of 10^{-16} watt. Energy supplied at this rate would take thirty thousand years to lift an object weighing one gram to a height of one centimeter!

548. Efficiency of luminous sources. The conversion of other forms of energy into light by artificial sources of illumination is extremely wasteful, because the greater portion of the energy supplied is converted into heat. The efficiency of this process is the ratio of the mechanical equivalent of the light emitted to the power consumed by the source. This power in artificial sources is supplied either by combustion, as in oil or gas lamps, or by a current of electricity. In the latter case it is easily measured, while the equivalent power output is obtained as described in the preceding article.

If the mechanical equivalent of white light were definitely 0.01 watt per lumen, the calculation of the efficiency of an electric light from its measured candle power and electric power input would be extremely simple. Thus if it has C candle power, it radiates a flux of

[†] If we assume the pupil of the eye to be dilated to a diameter of 4 mm, the conical intensity of 10^{-13} lumen perceived would result from a standard candle at a distance of 14 miles.

$4\pi C$ lumens. If this is multiplied by 0.01, the mechanical equivalent of its luminous output is $0.04\pi C$, and if it consumes P watts, the efficiency is given by

$$e = 0.04\pi C/P. \quad (1)$$

A modern gas-filled tungsten incandescent lamp of moderate power requires about 0.9 *watt per candle power* (a value often erroneously referred to as its "efficiency"). Therefore the ratio $C:P$ equals $1/0.9 = 1.11$ candle per watt, or $4\pi \times 1.11 = 14$ lumens per watt, approximately. The white-light efficiency of such a lamp (assuming 0.01 as the mechanical equivalent) is $14 \times 0.01 = 14$ per cent. The ordinary carbon arc has an efficiency of 11.8 per cent, while the new sodium arc has a net luminous efficiency of 60 lumens per watt, or 60 per cent of the electrical energy consumed.

This method of calculating the efficiency is open to two serious objections. These are, first, the uncertainty of the value of the mechanical equivalent of white light, and second, the fact that white light is not as efficient an illuminant as light concentrated in the yellow-green. Therefore, even if the value 0.01 watt per lumen were exact, it still involves a comparison between an actual source and one whose efficiency is not 100 per cent. To obtain this monochromatic efficiency we should take 0.00161 watt per lumen as the mechanical equivalent, and then

$$e = 0.00644\pi C/P. \quad (2)$$

This amounts to comparing the output of a lamp with that of an ideal source, which, consuming energy at the same rate, concentrates it in the brightest part of the spectrum. The efficiency calculated by equation (1) is 6.21 times as high as when (2) is used. Therefore a tungsten lamp measured by this more exacting standard has an efficiency of only 2.3 per cent. In a similar manner a carbon arc light has a monochromatic efficiency of about 1.9 per cent, a mercury arc of 6.8 per cent and a sodium arc of 9.7 per cent. But a glowworm exceeds 90 per cent.

It is evident that we are wasting an appalling amount of energy in our artificial illumination, and it is to be hoped that we shall some day know how to produce the "cold light" of the firefly and glowworm. At present the nearest approach to artificial cold light of real intensity is that produced by the oxidation of some organic compound. The most brilliant of these combinations is when luminol (3 aminophthalhydrazide) is oxidized in an alkaline solution with potassium ferri-cyanide and 3 per cent hydrogen peroxide.† The result is a pale

† *Journal of Chemical Education*, Vol. 2, p. 142, 1934.

bluish light strong enough to read by. But unfortunately the effect lasts only a short time, the light fading gradually until at the end of about a minute it is no longer more than a faint glow.

SUPPLEMENTARY READING

Hardy and Perrin, *The Principles of Optics* (Chap. 9), McGraw-Hill, 1932.

H. B. Lemon, *Cosmic Rays Thus Far*, W. W. Norton, 1936.

R. W. Wood, *Physical Optics* (Chap. 20), Macmillan, 1934.

F. A. Lindemann, *The Physical Significance of the Quantum Theory*, Clarendon Press, 1932.

PROBLEMS

1. Calculate the wave length of maximum energy emitted by a "black body" when heated to the melting point of platinum, 1755°C . *Ans.* 1.42μ .

2. What is the wave length of maximum energy emitted by platinum at its own melting point? *Ans.* 1.30μ .

3. What is the black-body temperature of an arc light which radiates energy having its maximum at a wave length of 0.7μ , just below the visible red? *Ans.* 3848°C .

4. Calculate the energy of a quantum of sodium light, also the mass and momentum of the quantum. *Ans.* 3.372×10^{-12} erg; 3.75×10^{-33} g; 1.126×10^{-22} dyne-second.

5. Calculate the wave length of the fourth line in the Paschen series of hydrogen in air. *Ans.* 1.005μ .

*6. A line in the Balmer series of hydrogen is found to have a wave length in vacuo of 4.34×10^{-5} cm. Substituting this figure for λ in the equation, show that the line is the third in the series.

CHAPTER 42

Optical Phenomena in Nature

549. The mirage. When the successive layers of air just above the surface of the earth differ in temperature, their density also differs, and light passing through them travels at different speeds according to the level above the earth's surface. This results in making the wave front of light from a distant source slightly curved in a vertical plane, as by the action of a cylindrical lens, even when the object is so far off that the curvature would otherwise be negligible. The most usual case is when there is a layer of heated air lying close to the earth, with cooler air above. Although this is an unstable condition and therefore variable, it is observed over hot desert sands and over the relatively warm water of the ocean in the early morning, after the upper layers of air have been cooled by radiation during the night.

The result of this unequal temperature has been explained by Hastings† in the following ingenious manner: The wave front in the

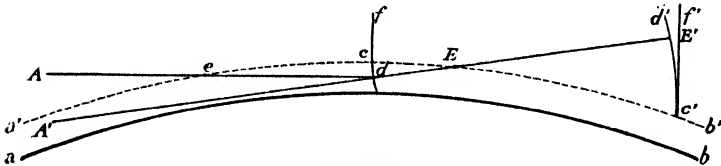


Fig. 139.

homogeneous layer high up gradually becomes plane, while that in the progressively warmer layers below becomes concave forward. Thus in Fig. 139, ab represents the curved surface of the earth, enormously exaggerated, with a layer of warm air $a'b'$ lying above it. As light from the object A enters this layer at e , the lower portion cd of the wave front becomes slightly concave, while the upper portion cf is almost plane. This is because the light travels faster through the less dense air at progressively lower levels.

† Charles S. Hastings, *New Methods in Geometrical Optics* (Appendix C), Macmillan, 1927.

An eye at E would see the object A , owing to the cylindrical portion of the wave front cd , at a lower point A' , and slightly magnified, as by a converging lens held close to the eye. But as this is a cylindrical wave front having a horizontal axis, the magnification is only vertical. At a still greater distance, with the eye at E' above the layer of heated air, there are two wave surfaces, or *sheets*. These are the undisturbed plane wave $c'f'$, and a slightly convex cylindrical surface $c'd'$ due to the reversal of the concave cylinder, just as a converging beam of light diverges after passing through a focus. The plane wave front has one erect virtual image of the object at A , but the convex front $c'd'$ has a virtual, inverted, and slightly magnified image at A' . The effect is of an object and its inverted mirror image, as a tree reflected in a lake. The illusion of water is still further accentuated, because the sky behind A is also brought down to A' by the curved portion of the wave front, so that the traveler in the desert seems to see a distant palm tree reflected in a pool of water, where there are only the tree and heated sand.

The same phenomenon in miniature is seen across the rounded surface of a concrete road at the crest of a hill, the curved hilltop replacing the curvature of the earth shown in Fig. 139. Here the image of the sky, owing to the cylindrical curvature of the wave front, is formed below the horizon defined by the hilltop, giving an illusion of pools of water. This is a sight familiar to motorists on hot summer days, but it is possible only when the crest of the hill forms or nearly forms an horizon for the observer.

The other kind of mirage is due to a cold layer of air over cold water with a warmer layer above. This is common in the early

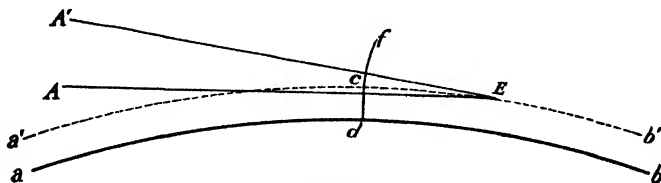


Fig. 140.

autumn in northern waters. In this case the lower layer is of fairly even density, while the upper one is progressively warmer and lighter. Hence the cylindrical wave front now forms *above* the plane one, and an eye at E , Fig. 140, sees the object A at A' , erect, magnified, and higher up. The author has seen cliffs on the northern shore of Lake Superior (a very cold body of water) magnified in this way to

several times their proper height. Distant boats appear abnormally high out of water, but of course no broader than without the mirage. Among seamen, this effect is known as *looming*.

At a still greater distance than that indicated in the diagram, when *cf* has reversed, an inverted image at A' may be seen. The plane portion of the wave front then yields an erect image of A , and the now convex wave front forms an inverted and vertically magnified image above it. In this way ships apparently sailing upside down in the sky are sometimes seen.

Unless the reversal of curvature (an extremely rare occurrence) has taken place, the eye must be above the cold layer in order to perceive the image formed by the cylindrical sheet *cf*. The image may be that of an object below the horizon, visible if the eye is above $a'b'$, and invisible if it is below. Thus a few inches up or down are enough to determine the visibility or otherwise of the object below the horizon. It is really a startling experience to lower one's head a foot or two and see a distant island, which was made visible by the mirage, disappear completely from view.

550. The form of the rainbow. When a ray of sunlight falls upon a drop of water it is refracted as by a lens, and emerges as part of a converging beam. But before it passes out of the drop it may be internally reflected one or more times, and then, because of the longer internal path, the resulting beam tends in general to be divergent. In Fig. 141 a horizontal ray of light is seen entering a spherical drop of water at p . A portion is internally reflected at a and again at p' , but a part emerges toward p'' , as indicated. Its direction has thus been altered by a very large angle approaching complete reversal.

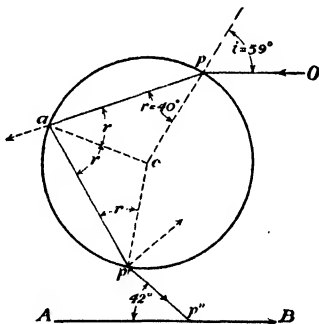


Fig. 141.

This angle of deviation is found to be a minimum when the angle of incidence is 59° . Then the angle of refraction is 40° , when n for water is 1.333. It is now easy to show by the trigonometric relations indicated in the figure that the emergent light has been deviated by 138° , so that the ray $p'p''$ makes an angle of 42° with the line AB drawn toward the sun.

It can also be shown that in the particular case assumed above, a narrow pencil of rays around Op as an axis is less diverged after emer-

gence than all other horizontal rays, which necessarily follow different paths through the drop. These, on emergence after one or more internal reflections, are so divergent as to be rapidly dissipated. But a pencil of light which follows the path pap' is not scattered in this way, and an observer at p'' would see the drop as a bright point of light if its angular relations to him and the source were defined by the angle of 42° , as in Fig. 141.

In order to understand the shape of the rainbow, consider a cone of circular section, shown in Fig. 142, whose base $abcd$ is perpendicular

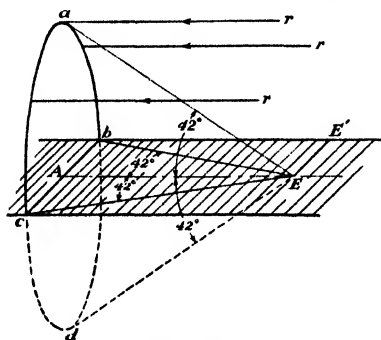


Fig. 142.

to the rays of light r . Its elements aE , bE , and so on, which meet at its vertex E , are inclined at 42° to its axis AE , where A is opposite the sun and is called the **antisolar point**. Any point on the surface of such a cone meets the requirements of the drop in Fig. 141 with respect to an observer at E . Then light from the sun, evidently assumed to be on the horizon, will reach the eye by the route of minimum deviation.

So any raindrop that happens to be anywhere on the surface of the cone is seen as a luminous point, and if a shower is falling in the region to the left of E , the luminous points formed by the drops as they pass through the conical surface appear as the arc of a circle, although actually they form a conical sheet which is seen edgewise.

A given drop meets the requirements of the rainbow for a given observer only for an instant, but as there are many other drops to take its place, the rainbow seems continuous, although the individual drops forming it are constantly changing. Moreover, another observer at E' sees another bow formed by a different set of drops, although his cone has an axis parallel to that of the observer at E , and the two cones and rainbows are otherwise exactly similar. Hence the futility of seeking the fabulous pot of gold at the foot of a rainbow!

When the sun is above the horizon, the cone of minimum deviation is tilted, with its axis always a continuation of the line between the sun and the observer. Thus with the sun at an elevation of an angle α , the visible bow is only a portion of the circumference of the circle cut off by the earth's surface, as shown in Fig. 143. When $\alpha = 42^\circ$, no bow appears above the xy plane, which explains why rainbows

are never seen when the sun is high above the horizon, but only in the early morning or late afternoon, except in high latitudes.

It is possible to see the entire circle of the bow, however, when standing on a steep elevation and looking off into a nearby shower of rain, or spray from a waterfall, provided the sun is shining behind the observer and is not too high.

551. The colors of the rainbow. The angle of the cone where the rainbow is formed, as stated in the last paragraph, is 42° , but only for red light, as the value of the refractive index

(1.333) used in Article 550 is for that color. The other colors of the solar spectrum are more deviated by a drop of water, and the total minimum deviation for violet, after one internal reflection, is 140° , instead of 138° , as in the case of red light. Therefore the angle of the cone for violet is 40° instead of 42° . Thus there is a series of conical sheets lying one inside the other, and bounded by the 42° and 40° cones, as in Fig. 144. This results in a series of circles or parts of circles as seen by the eye at E , with red on top and graduating through the colors of the spectrum to violet at the bottom.

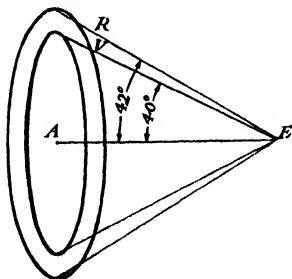


Fig. 144.

552. Secondary bows. These are formed after two internal reflections, as shown in Fig. 145. Here the sun's light, if horizontal, must enter near the bottom of the drop in order to reach the observer, and on emergence it makes an angle varying from 51° for red to 54° for violet light. In this way a second set of cones is constructed, having larger angles than the primary set, and with violet lying above the red. This secondary bow is necessarily fainter than the primary, because of the increased loss of light at the second

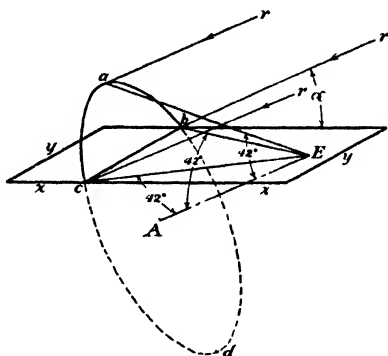


Fig. 143.

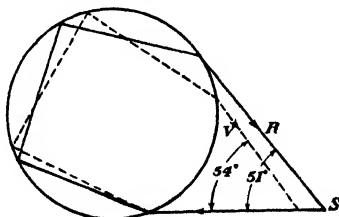


Fig. 145.

internal reflection, which is only partial. But it is a little broader, since the difference between its limiting angles is about 3° , instead of 2° , and the colors are in reversed order.

Bows of still higher orders are possible, but the third and fourth orders are only about 50° from the sun, and are therefore much too faint to be seen against such a brilliant background as the brightest part of the sky. The fifth-order bow, whose angle with respect to the antisolar point is 54° , is too faint to be seen even at this favorable angle, because of the progressive loss of light involved in five internal reflections.

553. Coronas and halos. Sometimes when the sun shines through a fog we see a colored band surrounding it, like a rainbow of small angular aperture, but with less well-defined colors. Such a **corona** is even more common around the moon when there is mist in the air, and its apparent diameter varies according to the height of the mist. These coronas, popularly known as rings around the sun or moon, are due to diffraction by the particles of suspended moisture. They are easily reproduced on a small scale by holding a piece of glass, lightly fogged with steam, between the eye and a point source of light.

Halos are formed in a manner similar to rainbows, but in this case small ice crystals in the upper atmosphere take the place of the rain-

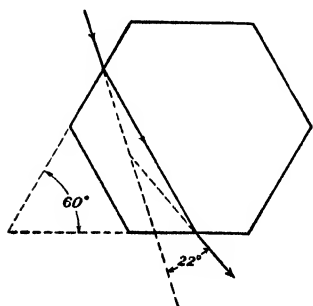


Fig. 146.

drops. These are usually right-hexagonal crystals having 60° angles between alternate faces, as in Fig. 146, and these faces then constitute a truncated prism having a refracting angle of 60° . The angle of minimum deviation can be calculated from the prism formula, and is found to be 22° when n is set equal to 1.31, the index of refraction of ice. Therefore when myriads of such tiny crystals are falling between the observer and the sun or moon, some of them

will always be correctly oriented to bend the light to the observer's eye without the dissipation due to divergence, which would occur at other angles. This creates a ring of prismatic colors with red nearest the sun or moon, a bluish hue outside, and having an angular radius of 22° .

Another halo, caused by the 90° prisms which the lateral faces of the crystal make with its bases, is also possible. In this case the angle of minimum deviation is 46° , and the resulting halo, in consequence, has a 46° radius around the sun or moon. This is much rarer than

the smaller type, but is occasionally very brilliant when the crystals are sufficiently numerous.

Sundogs, or parhelia, are vertical areas usually seen on either side of the setting sun at 22° angular distance. They are caused by hexagonal ice crystals falling with their axes vertical, so that the 60° prism described above becomes effective only when the crystal has the same altitude as the sun.

There are various other rare phenomena of this sort, involving predominant orientations of the crystal in unusual positions, such as when the majority are falling with their axes horizontal, but enough has been said to account in a general way for these interesting natural spectacles.

554. Color of the sky. In order to understand why the sky is blue, it is necessary to investigate the phenomenon discussed in Article 517 concerning the diffraction of light by minute particles. The intensity of the light scattered sidewise in this way varies inversely as the fourth power of the wave length, so that the shorter waves toward the violet end of the spectrum are sent out at right angles to the beam much more freely than the longer waves near the red end. This results in a sorting of the light, so that a pencil of white light sent through a suspension of fine particles appears bluish when seen sidewise, and reddish when viewed toward the source, because of the sorting out of the green, blue, and violet.

The red color of sunset is due to a similar sorting in which minute dust particles near the earth play an important part. When the sun is low on the horizon, it shines through a much thicker layer of air than when it is overhead, and consequently the shorter waves have a better chance of being scattered.

To account for the blue of the sky we must first suppose that the molecules of air act in a manner similar to fine dust particles with respect to scattering, and also that the effect is a double one. Thus a beam of light, if it reaches us from the portions of the sky away from the sun where the blue is strongest, must have been sent out from the sun at some angle α to the line of sight, which would give the beam a component H normal to that line, as shown in Fig. 147, where the observer is supposed to be looking straight up. The scattered light of these transverse components, seen from the ground without further

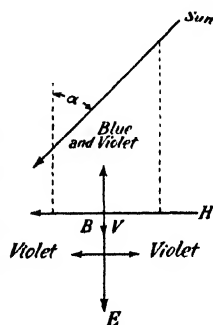


Fig. 147.

modification, would appear blue-violet instead of blue, because of the more effective scattering of the shortest waves of the spectrum; and indeed this color has been observed from balloons at a very great altitude. But near the surface of the earth the light scattered by the upper strata of air is again scattered in passing downward through the denser atmosphere, and during this second scattering the violet is thrown out sidewise from the descending beam, which reaches the eye at *E* with a residue of blue. This blue light is partly polarized, as explained in Article 517, and can be observed by looking at the sky through a Nicol's prism. The sky then appears darker if the Nicol is "crossed" with the plane of vibration of the light.

SUPPLEMENTARY READING

R. W. Wood, *Physical Optics* (Chap. 11), Macmillan, 1934.

J. Preston, *The Theory of Light* (Chap. 20), Fifth Edition, Macmillan, 1928.

W. J. Humphreys, *Physics of the Air*, Second Edition, McGraw-Hill, 1929.

PART V
ELECTRICITY AND MAGNETISM

CHAPTER 43

Magnetism

555. Nature of magnets. The property called magnetism, by which certain metallic substances attract others at a distance, was known to the ancients. They found in Magnesia, Asia Minor, a certain mineral, composed of the oxides of iron FeO and Fe_2O_3 , which possessed natural *magnetism*. The name was probably derived from that of the region in which the ore was first found. This ore, known as magnetite, together with the weaker magnetic pyrites, constitutes the only natural magnet known. Magnetite is also called "lodestone" (leading stone), and was first scientifically investigated by William Gilbert, an English physician, whose famous treatise "De Magnete" appeared in the year 1600.

If a piece of magnetite is dipped in iron filings, they cling to it, especially at certain points where the magnetism seems to be concentrated. These points are called the **poles** of the magnet. A bar of steel stroked with a lodestone acquires poles, a fact known to the Chinese, who invented the mariner's compass. Such a bar has ordinarily but two poles situated near its ends, as may be seen from the concentration of iron filings there, when they are sprinkled over it. A magnetized bar, if pivoted like a compass needle, always lies in the north and south magnetic meridian, and the same end always points north. This shows at once that the poles are different in kind, and the difference is still further exhibited by the action of two pivoted needles brought near each other. In this case, two ends of the same kind will be found to repel each other, while two of opposite kind attract. Thus we derive the familiar law, *like poles repel and unlike poles attract*. The north-seeking poles are called north poles in ordinary practice, and the south-seeking poles, south poles, though it should be noted that these terms are misnomers, because the poles of a magnet are of opposite kind from the poles of the earth for which they are named. They are also often referred to as positive and negative poles.

556. Coulomb's law of attraction. An ideal pole is one so far removed from its mate as to be practically beyond its influence. It

must also be of such small dimensions that it may be regarded as a point. In 1785, C. A. de Coulomb, a French physicist, investigated the properties of magnetic poles, using long and slender magnetized needles in an apparatus similar to that described in Article 580. In this way he realized approximately ideal poles and formulated the laws of their mutual action. These may be combined in a single statement known as Coulomb's law: *The force of attraction (or repulsion) between two poles varies directly as the product of their strengths, and inversely as the square of the distance between them.* Expressed symbolically, this law may be stated as

$$F \propto \frac{mm'}{r^2}, \quad (1)$$

where m and m' are the strengths of the two poles, and r is the distance between them. If both poles are positive or both negative, F is positive, and we know from experiment that the poles repel each other. If they are of opposite sign, F is negative, and the poles attract each other.

557. The unit pole. Coulomb's law tells us how we may define a unit pole. In order to do so we must specify the medium between the poles, because it is found that the force depends upon the medium. It is therefore agreed to assume that the action takes place in a vacuum. Then we agree that m shall be so defined that the force not only varies as mm'/r^2 , but shall be numerically equal to mm'/r^2 . Therefore if F and r are unity, mm' must be unity also, and further, if m and m' are equal, each is a unit pole, whose strength may be denoted by e.m.u., an abbreviation of *electromagnetic unit*. Thus we arrive at the definition, that *a unit pole is one which at a centimeter's distance in a vacuum acts with a force of one dyne upon another pole of equal strength.* Our unit having been thus defined, Coulomb's law becomes

$$F = \frac{mm'}{r^2}. \quad (2)$$

This is rigorously true, by definition, in a vacuum; but it is also so nearly true in air that for most purposes we may use equation (2) as if it were exact when air surrounds the poles.

Equation (2) also shows what is meant by a pole whose strength is greater than unity. If $m = 1$ and $m' = 2$, the force is twice as great as when both are unit poles. If both poles have a strength of two units, the force is four times as great as between unit poles, and so on.

The dimensions of a unit pole are obtained from Coulomb's law as follows: Since the dimensions of m and m' are the same, we may write

$$[F] = [m^2 L^{-2}].$$

But in general

$$[F] = [MLT^{-2}].$$

$$\therefore [m] = [M^{\frac{1}{2}} L^{\frac{1}{2}} T^{-1}].$$

558. Magnetic field. The region around a magnet, in which its effect is perceptible, is known as its **field**. The field is stronger near the poles, and becomes weaker as we recede from them. It can be

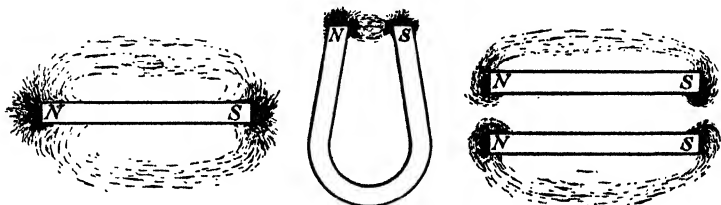


Fig. 1.

examined by sprinkling iron filings on a piece of paper held over the magnet. They then form themselves in such patterns as are indicated in Fig. 1. It will be seen that the filings arrange themselves in lines, by placing themselves end to end under the influence of poles induced in them, thus suggesting the varying direction of the field, as well as its varying intensity indicated by the density of their grouping.

The direction of these lines can be studied still better with the aid of a small compass placed in the same horizontal plane as the magnet. The effect on the compass needle is shown in Fig. 2. Let sn represent the needle and NS the magnet. Four forces are acting, involving one of repulsion and one of attraction at each end. The resultants of these pairs act in part as a couple tending

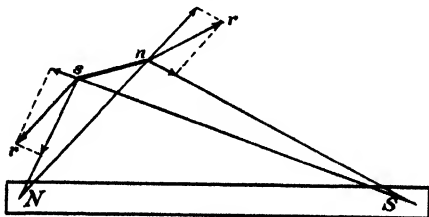
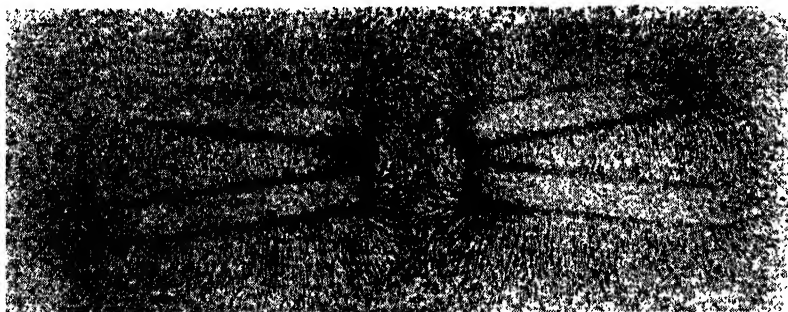


Fig. 2.

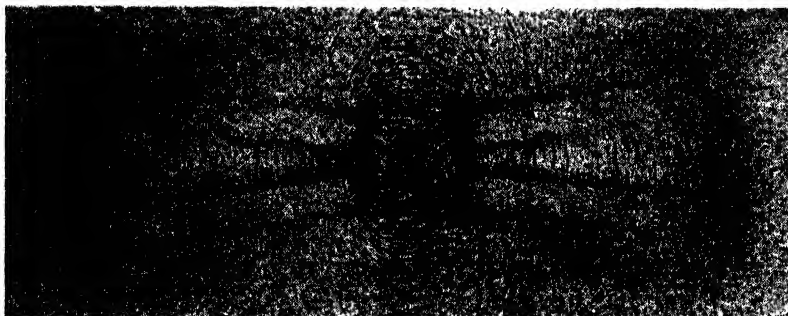
to turn the needle about its axis in a counterclockwise direction, until it sets itself, as shown in Fig. 3, when the turning moment vanishes. The needle is then tangent to the direction of the field at its axis. In this way, by placing the needle at successive positions,

the field may be mapped with considerable accuracy. It should be remembered, however, that the earth's field is always present, so that the field indicated by the compass is the resultant of the combined fields of the earth and the magnet.

On careful examination of Fig. 3, it will be seen that the resultant force acting on the south pole of the needle must be greater than that



(a)



(b)

Plate 15.

(a) Photograph of iron filings sprinkled on cloth stretched over horseshoe magnets, showing field with like poles opposed (repulsion). (b) Photograph of iron filings showing field of magnets with unlike poles opposed (attraction).

acting at the other end. This is because it is nearer the north pole of the magnet, and the attraction vector, to which most of the resultant is due, is slightly larger than the repulsion vector at the other and more distant end of the needle. The result is an unbalanced force tending to *translate* the needle toward the nearest pole of the magnet. Moreover, the two resultants do not lie quite in the same straight line, so that there is a small unbalanced force acting at right angles to the needle and tending to draw it sidewise. If the needle is free, as when

mounted on a floating cork, it moves both longitudinally and transversely toward the attracting magnet. A similar motion also takes place when an unmagnetized piece of iron is near a magnet, or in general, wherever there is a nonuniform magnetic field. Then motion results in the direction of increasing field intensity.

If the length of the compass needle in Fig. 3 is made very small, the difference between the two resultant forces becomes negligible, they are oppositely directed, and the needle does not move toward the magnet. This would be the case when any needle is acted on by the earth's field, for it is vanishingly small in comparison with the distances to the magnetic poles of the earth. Thus, if a floating magnet were acted on by the earth only, we should have rotation, but no translation. The same is true in any uniform field, as will be explained later.

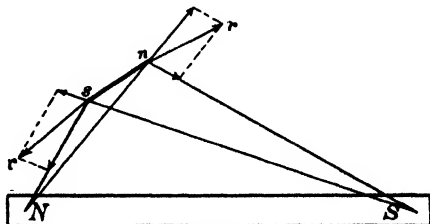


Fig. 3.

559. Field intensity. In order to measure the effect of a magnet on the region surrounding it, we make use of a quantity called **field strength**, or **field intensity**. It depends upon the position and strength of the magnet, and is defined as *the force in dynes acting upon a unit north pole at a specified point.*† Since the force has a definite direction, field strength is a vector quantity. It is denoted by H , and may be calculated by dividing the force on an ideal pole m' by its pole strength, or

$$H = \frac{F}{m'}.$$

The unit of field intensity is now called the **oersted**.‡ The oersted is defined as the intensity at a given point in a field at which the field would act with the force of one dyne upon a unit pole placed there. Or 1 oersted = 1 dyne/1 pole. Its dimensions are those of a force divided by the dimensions of pole strength, and are given by

$$\begin{aligned} [H] &= [MLT^{-2}/M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}] \\ &= [M^{\frac{1}{2}}L^{-\frac{1}{2}}T^{-1}]. \end{aligned}$$

† This is on the assumption that the field is undisturbed by the presence of the pole.

‡ Formerly called the *gauss*. Renamed after the Danish physicist H. C. Oersted (1777-1851).

The value of H at any point in a magnetic field is the vector sum of all the forces which act upon a unit pole at that point. Thus, in Fig. 4, the field at p is due to both poles of the magnet. Each component is the actual force on a unit pole at p , and these are then added vectorially, as indicated.

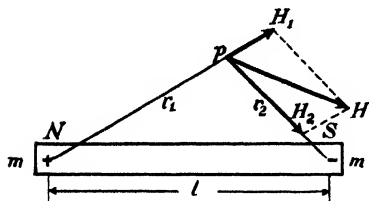


Fig. 4.

We may calculate H at a given point as follows: The force due to the pole N of strength m is given by Coulomb's law. Then since $H = F/m'$ we have

$$H_1 = \frac{F}{m'} = \frac{mm'}{r_1^2 m'} = \frac{m}{r_1^2} \quad (1)$$

which is the force on a unit pole ($m' = 1$) at a distance r_1 from N . Similarly, $H_2 = m/r_2^2$. The resultant field H is the vector sum of H_1 and H_2 , and may be obtained graphically or by trigonometry if l , r_1 , and r_2 are known. The value of H at p would really be altered by the presence of a magnetized needle there, but this fact does not affect the results obtained by calculation, which give field intensity and its direction independent of any actual pole at p .

560. Lines of force. It is a great help in studying magnetic fields to have some means of picturing them to the eye, and in a manner useful in making calculations. We saw in Article 558 that iron filings, in a magnetic field, map lines indicating the direction in which a small magnet tends to lie. They also tell us by their distribution where the field is intense and where it is weak. Instead of using iron filings, we may explore the field with a small compass and plot the lines on a sheet of paper. These are called **lines of force**, and are defined as *purely imaginary lines or curves which start on a positive or north-seeking pole and end on a negative or south-seeking pole. Their direction at any point in their course is that of the field H , or the direction in which a single positive pole would travel if placed there.*

Even as specified above, lines of force, by crowding together at some points and spreading out at others, would give us information only as to the field's direction and where it is stronger and where weaker, but they would tell us nothing of the actual value of the field's strength. To make this possible, let us draw a definite number of lines of force starting out symmetrically in all directions from each unit pole when considered all by itself. As these lines spread out radially, their number crossing a unit area normal to their direction must decrease with increasing distance, like everything else

which spreads out uniformly in three-dimensional space. That is, like luminous flux, their density must vary inversely as the square of the distance from the pole. But H decreases in the same way, so that the number of lines of force crossing a unit area normal to their direction is a correct measure of the field strength.

In deciding how many lines we shall imagine as starting from a unit pole, we may measure the field strength one centimeter away from the unit pole. From equation (1) in the last article, $H = m/r^2$, and as m and r now both have unit value, H equals one oersted. But the entire region in which $r = 1$ is a sphere surrounding the unit pole, and such a sphere has an area of 4π cm². So if we decide to have 4π lines radiate from an isolated unit pole, one line will cut across every square centimeter of the surrounding unit sphere. Thus in a field where $H = 1$, there will be one line of force for each square

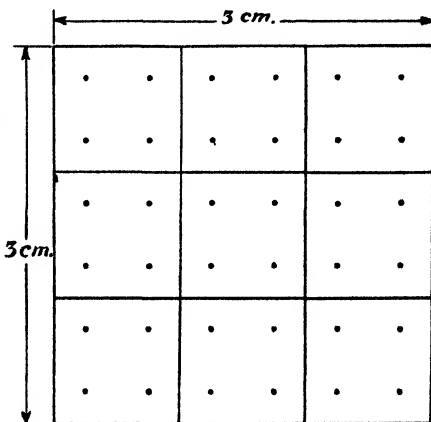


Fig. 5.

centimeter normal to the field. If there are four lines cutting through each square centimeter, as indicated by the dots in Fig. 5, then $H = 4$, and so on. As this is a very convenient measure, it is usually assumed that 4π lines of force start from a unit positive pole and end on a unit negative pole.

561. Magnetic moment. In many calculations the product of the pole strength of a magnet and the distance between the poles must be found. But neither the position of the poles nor their strength can be accurately determined, so that a new unit has been adopted called **magnetic moment** (M). It is defined as the product just referred to, or ml , and is readily measured. Its unit is a unit-pole-cm. The need for such a unit is seen when we consider the turning moment on a suspended bar magnet in a uniform field of intensity H , as shown in Fig. 6. The two forces acting on the poles are each equal to Hm . The turning moment due to each is obviously $Hm \frac{l}{2} \sin \alpha$, where l is the distance between the poles and α is the angle between field and

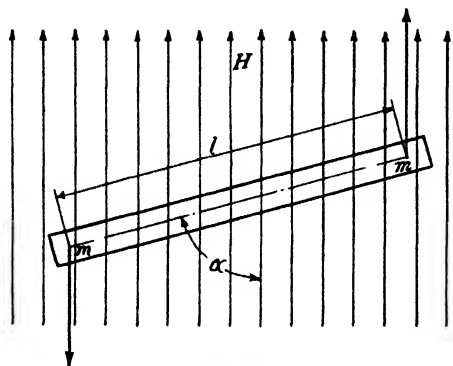


Fig. 6.

magnet. Therefore the total torque L is given by

$$\begin{aligned} L &= 2Hm \frac{l}{2} \sin \alpha \\ &= Hml \sin \alpha \\ &= HM \sin \alpha, \end{aligned}$$

so that if M and H are known, the torque for any angle may be calculated. It is well to note that L vanishes when the axis of

the magnet is parallel to the field, and is a maximum when it is perpendicular. In the latter case $L = HM$, and if H is unity, $M = L$. Thus magnetic moment may be defined as the torque exerted by a unit field on a magnet placed at right angles to the lines of force.

562. Magnetic dipoles. If the magnet suspended in a field, as described in Article 561, is cut in two, each half will be found to be a complete magnet with a north- and a south-seeking pole. Each part will be found to have half the magnetic moment of the whole magnet. This means that the poles must be as strong as before, because if ml is half what it was, and l is halved, m is unaltered. If we then cut each of these two magnets into two equal parts, we would have four complete magnets, each having one quarter of the moment of the original bar. No matter how far we carry on this subdivision, we always obtain magnets with two poles each, and the sum of their individual moments is always equal to the moment of the undivided magnet.

The facts described above led Wilhelm Weber, a German physicist (1804–1890), early in the last century to propose the theory that substances which could be magnetized were made up of minute molecular magnets. These elementary **dipoles** were supposed to have their axes pointing at random in all directions until the material was magnetized. Then they were thought to fall into line with their *axes* parallel to each other and to the axis of the magnet. For instance, if an iron bar consists of molecular dipoles oriented wholly at random, their fields would neutralize each other, and the bar as a whole would be unmagnetized. But if the dipoles were lined up with like poles all pointing the same way, their individual fields would help each other, and the bar as a whole would become a magnet with

free poles at its ends. A chain of dipoles reaching from one end of the magnet to the other may be represented by N -s, n -s, n -s, n -S, where the capital letters represent the *free* poles at the ends, all other poles being tied together by mutual attraction. Weber's hypothesis seemed justified by a number of facts, among them being a change in the dimensions of an iron body when it becomes magnetized. This was first observed by Joule, and is called **magnetostriction**.

563. Intensity of magnetization. In order to measure the strength of a magnet, a quantity known as **intensity of magnetization** is employed. This quantity is defined as the magnetic moment per unit volume, and is denoted by the letter \mathcal{Q} . Thus

$$\mathcal{Q} = \frac{M}{v}, \quad (1)$$

where $M = ml$, as already noted, and v is the volume of the magnet. If the magnet is a bar of uniform sectional area a , $v = al$, and we may write

$$\mathcal{Q} = \frac{ml}{al},$$

or

$$\mathcal{Q} = \frac{m}{a}. \quad (2)$$

Thus the unit of magnetic intensity is *unit pole per cm²*. But near a pole of strength m , $H = m/r^2$, so that unit H may also be defined as unit pole per cm². Therefore H and \mathcal{Q} measure the same kind of thing and have the same dimensions.

In order to form a clearer picture of this important quantity \mathcal{Q} , we may imagine the ends of a long bar magnet, cut apart at its center, brought face to face, N opposite S , across a narrow air gap. If each pole has a strength of m e.m.u., we may picture the two ends as each having m unit poles distributed evenly over its section area a . Now, as we have just seen, $\mathcal{Q} = m/a$; therefore \mathcal{Q} measures the *surface density of pole strength*.

From the foregoing conception of \mathcal{Q} , a very important relation between \mathcal{Q} and H may be derived as follows: Since the two poles are supposed to be close together, the field between them is practically uniform, and $H = 4\pi m/a$. But $m/a = \mathcal{Q}$; therefore, within the gap,

$$H = 4\pi\mathcal{Q}. \quad (3)$$

This way of calculating H from \mathcal{Q} may also be used to give approximate values of H at points very close to the ends of a single bar mag-

net when its magnetic intensity is known. Or \mathcal{Q} may be found at such points from H . Numerical values of \mathcal{Q} may be very large. A bar of hardened steel may be magnetized to retain an intensity of over 900 (e.m.u.)/cm². Cast iron can retain an intensity of around 400. Natural magnets of magnetite are much weaker, but may be magnetized to retain an intensity of about 200 (e.m.u.)/cm².

564. Magnetizing iron. In Article 562, we saw that iron may be regarded as made up of minute magnets (dipoles) arranged wholly at

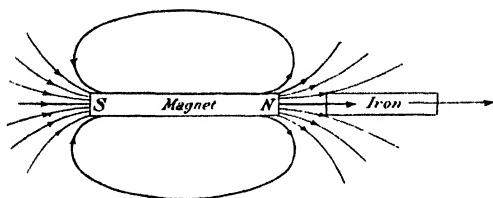


Fig. 7.

random. Let us now suppose that a bar of iron is brought into a field of magnetism, as shown in Fig. 7. Under the influence of the field, indicated by the curved lines, every little dipole in the iron

bar tends to line up along a line of force. Being more or less free to turn inside the iron, they do line up to a certain extent, and as a result of their united action, the iron develops **induced poles**, $N'S'$, at its ends, as shown in Fig. 8.

The process of inducing poles is really not as simple as just indicated. The field shown penetrating the iron bar is not as great as it was before the bar was introduced.

This is because of the effect of the induced poles. The end of the iron next the N pole of the magnet

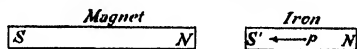


Fig. 8.

now has a lot of free S poles due to the S ends of the dipoles at that surface. At the surface farthest from the magnet, there are a lot of free N poles due to the N ends of the dipoles there. Hence the whole magnetic field at a point p (Fig. 8) in the iron will be due both to the original field of the magnet and to a new field acting against it, as indicated by the arrow. This field is due to the free poles induced in the iron, and is as usual directed from north to south polarity. It tends to weaken the field due to the magnet, and the actual magnetic moment induced in the iron is not as great as it would be if its own induced poles were not acting against the magnetizing tendency of the permanent magnet. This effect is called the *demagnetizing action of the free poles*. In the case of a short thick magnet, the intensity of this demagnetizing field is very strong. If the magnet is long and slender, the demagnetizing field approaches zero.

565. Permanent magnetism. The degree to which an iron bar becomes magnetized by a magnetic field is a very complicated problem. It depends upon the freedom with which the little dipoles in the iron may turn. That they are not absolutely free is shown by experiment, for if they were, the weakest field would line them up completely, producing *saturation*, as such a state of magnetism is called, and this does not happen. Another evidence that the dipoles must be restricted from turning freely is the fact that once lined up they tend to remain more or less fixed in that condition. If this were not so, permanent magnets would be impossible, because their internal fields, due to their own poles, tend to swing the dipoles around in a direction opposite to that in which they are pointing.

Suppose for a moment that the dipoles in a magnet were quite free to turn in a field due to the poles created by their united action at the ends of the bar. Then they would start swinging around to point the other way, and as more and more lined up in the reverse direction, the free poles at the ends would become steadily weaker. This weakening of the free poles would in turn result in a decrease of the demagnetizing action, so that the process of demagnetization would proceed, but with less and less vigor. Thus the magnetization would gradually settle down to zero, and the magnet would become just a piece of unmagnetized iron. The fact that this does not usually happen must be due to a lack of freedom of the dipoles to turn unless strong fields act upon them.

Different kinds of iron and steel behave very differently in retaining magnetism. If we bring a piece of very soft iron near a magnet, it is magnetized by induction, but when we take it away it becomes almost demagnetized by the action of its own poles, as just described. If a piece of hard steel is brought near a magnet, it too is magnetized by induction. But when we take it away, its dipoles do not turn readily under the demagnetizing action of its own poles, since this action is much weaker than the field which produced its poles. Therefore the bar remains permanently magnetized.

566. Ferromagnetism. Besides iron, a few other substances develop a strong intensity of magnetization when put in moderate magnetic fields. As their behavior is similar to that of iron, they are called **ferromagnetic** (from Latin *ferrum* = *iron*). Steel, which is mostly iron, is of course ferromagnetic. In addition, the elements nickel and cobalt are strongly magnetic but less so than iron. These three form the so-called *ferromagnetic group* of elements. Quite recently the rare metal gadolinium has also been shown to be ferro-

magnetic when its temperature is below 16°C , its "Curie point" (see Article 568).

The ferromagnetic metals just named, with the exception of steel, are elements. But it is possible to make ferromagnetic alloys of non-ferromagnetic elements. The most famous of these are the Heusler alloys, which are intermetallic compounds such as Mn_3Sb_2 with another substance in solid, preferably concentrated, solution. This added substance may be antimony (Sb), manganese (Mn), or copper (Cu). Illustrations of such alloys are $\text{Mn}_3\text{Sb}_2 + \text{Sb}$, $\text{Mn}_2\text{Sb} + \text{Mn}$, and $\text{MnAl}_3 + \text{Cu}$.

Recently a new series of highly magnetic alloys has been developed for industrial purposes. These are generally alloys of two or of all three of the ferromagnetic group, iron, nickel, and cobalt. One of them, developed by the Bell Telephone Laboratories, is composed of 78.5 per cent of nickel and 21.5 per cent of iron. It belongs to a group of so-called "permalloys," because it has to a high degree the property known as *permeability*, to be defined in a later chapter. This means that it behaves as if its constituent dipoles were remarkably free, and so give it strong magnetic intensity in very weak fields, though \mathcal{I} disappears almost completely when the field is removed.

567. Para- and diamagnetism. Most substances are not ferromagnetic, but this does not mean that they are wholly indifferent to a magnetic field. If a sphere of aluminum is suspended by a thread near a magnet, it will be found to have a very slight tendency to go toward the strong part of the field. This behavior is like that which would be shown by a sphere of iron, except that the deflecting force is very much weaker with aluminum. To account for this we may suppose that aluminum, like iron, contains dipoles that are oriented by a field and thus create induced poles. Then by means of these poles the aluminum is attracted to the magnet which causes the field. Such weakly magnetic bodies are said to be **paramagnetic**. They were so named by Faraday,[†] because in the form of slender rods they tend, like iron, to set themselves parallel (from Greek *para* = *beside*) to the lines of force in a magnetic field.

In 1778, S. J. Brugmans (1763–1819), a Dutch physicist, observed that a piece of bismuth repelled either pole of a pivoted magnetic needle. A similar property of antimony was discovered in 1827, and in 1845, Faraday, using a very powerful electromagnet, found

[†] Michael Faraday (1791–1867), a celebrated English physicist and chemist, to whose extensive researches we owe much of our present knowledge of electromagnetic phenomena.

still other substances that tend to move away from strong into weaker magnetic fields. As slender rods of these substances tend to set themselves *across* the lines of force, Faraday named them **diamagnetic** (from Greek *dia* = *through*). Their behavior cannot be explained in terms of dipoles unless we are willing to suppose a very perverse variety of dipole which lines up in exactly the opposite direction to the way in which the magnetic forces are acting on it.

Although diamagnetic forces are very weak—too weak to have any practical applications—they are of the very greatest importance in giving us a deeper insight into the true nature of magnetism than we get from the simple dipole picture. But this cannot be discussed profitably until we have learned about the magnetic effects of electric currents in a later chapter.

568. Effect of temperature on magnetism. If a piece of very pure iron is placed in a very weak field, in which H is equal to 0.4 oersted, for instance, its permeability increases with the temperature to a maximum of 11,000 around 770°C . This is known as the “Curie point.” Beyond 770° , the permeability falls rapidly to almost unity at 800°C , and the iron becomes paramagnetic. If the field is stronger (say $H = 4$ oersteds), the permeability starts with a higher value, increases a little, and then falls more gradually to the same value at the same temperature. In strong fields there is no increase at all, but a gradual decline from a low value to nearly unity. These effects are shown in the curves of Fig. 9, where μ (Greek mu) represents permeability.

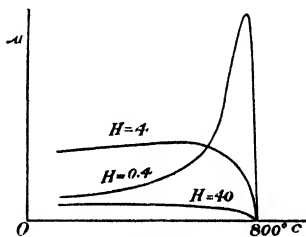
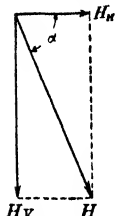


Fig. 9.

The critical temperature at which pure iron in a weak field reaches its maximum permeability corresponds to the so-called **point of recalescence**. This is seen when a piece of steel is allowed to cool rapidly from bright red heat. At a certain well-defined temperature, depending on the exact nature of the steel, it glows more brightly again, and then cools down normally. Here undoubtedly occurs a rearrangement of the molecular structure at a temperature below which the ferromagnetic properties of the substance are possible, and above which it acts like an ordinary paramagnetic body. In consequence of this phenomenon, a piece of iron heated to over 800°C is no longer attracted by a magnet, while a permanent magnet loses its magnetism and does not recover it again after cooling.

569. Terrestrial magnetism. The well-known action of a compass is due to the fact that the earth itself is a gigantic magnet, so that at all points near the earth's surface there is a field due to the magnetized matter in the earth's interior. How the earth came to be magnetized at all is not known. The compass points in the direction of the horizontal component of the earth's field. But this field at most places has a vertical component as well, so the compass needle is balanced to offset the very weak torque which tends more or less to deflect it from the horizontal plane.

To find the direction of the resultant field due to both the horizontal and vertical components, it is necessary to use an instrument called a **dip circle**. In this a magnet is pivoted on a horizontal axis so as to be free to turn in the vertical plane that passes through the direction in which a compass needle points, that is, in the "magnetic meridian." If the dipping needle is perfectly balanced *before* it is magnetized, then it is in neutral equilibrium under gravity. Now after being carefully magnetized, it will turn so as to point in the direction of the resultant field H and make an angle α with the horizontal, as indicated in Fig.



10. There the vector H represents the resultant intensity of the earth's field. H_H represents the horizontal and H_V the vertical component. The angle α is the angle of dip.

570. Declination and inclination. Since the earth's field at any point is a vector, we need to specify its direction and magnitude at that point. The angle which the horizontal component makes with the geographic north is called the **declination**, and the angle which H makes with the horizontal, as indicated by the dip circle, is called the **inclination**. These two angles suffice to tell us the direction of H . Finally we have to find the magnitude of H in order to describe it completely. This is, of course, the vector sum of the horizontal and vertical components shown in Fig. 10, or $H = \sqrt{H_H^2 + H_V^2}$. If H_H and the angle of dip α have been measured, we may calculate H from the obvious relation $H = H_H / \cos \alpha$.

The value of H varies considerably over the earth's surface, from a minimum of a little over 0.2 oersted to maxima of the order of 0.7 oersted. The horizontal component varies from a maximum of about 0.4 oersted near the equator to zero at the magnetic poles.

Magnetic surveys of the whole earth's surface have been made, so that much is known about the way in which the field is distributed, even though we do not know how the earth came to be magnetized as

it is. It is not to be expected that the field will be particularly simple, and it has been found that there are many local irregularities. But by and large the field is mainly such as would be produced by a huge permanent bar magnet in the earth's interior, with its axis making a small angle with that of the earth. The positive pole of this magnet is directed toward 72° S. latitude and 155° E. longitude (from Greenwich). The negative pole (toward which the north-seeking end of a compass points) lies on Boothia Peninsula, north by east from Hudson's Bay and almost due north of Omaha. Its latitude and longitude are about $70^\circ 5$ N. and 97° W. As these poles are not diametrically opposite each other, the main part of the field is somewhat lopsided with respect to the earth's center.

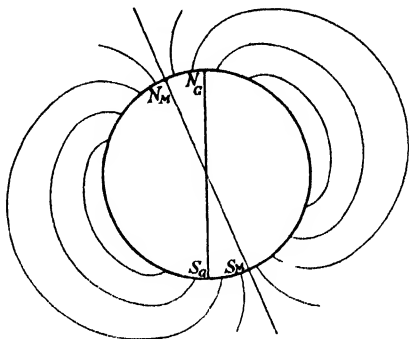


Fig. 11.

The general direction of the earth's lines of force about a section through its magnetic and geographic axes is shown in Fig. 11. From this diagram it is clear that a dipping needle would point straight up

and down at the magnetic poles, horizontally at the magnetic equator, and at angles varying between 90° and 0° at other points. From an examination of Fig. 12 we can understand why the compass needle in general does not point toward the true north. Points on a given curve in the diagram are points of equal declination, or **isogonic lines**. The heavy line indicates zero declination, although actually, between the

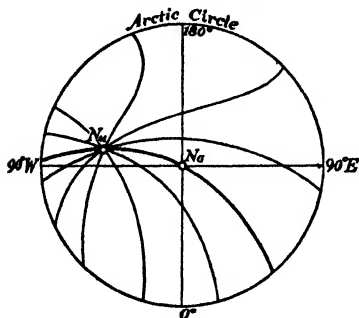


Fig. 12.

magnetic pole N_M and the geographic pole N_G the compass would point due south. East of that line it points west of north, and west of that line it points east of north. Since the meridian of zero declination (the heavy line) must form a closed curve passing through both magnetic and both geographic poles, it is evident that in circumnavigating

the earth, a ship's compass would twice point due north at points roughly diametrically opposite each other. In addition there is a closed loop of zero declination embracing Japan and parts of China and Siberia, known as the "Siberian Oval."

571. Variations of the earth's field. The various quantities we have been considering are constantly changing by small amounts. There are five types of such changes: A steady change known as secular; a daily variation; minute annual changes; others corresponding to the lunar month; and abrupt variations known as magnetic storms, which occur during periods of unusual solar activity, as shown in sun spots. These latter are apt to be associated with especially intense displays of the Aurora Borealis. The secular change in New England has averaged less than four minutes of arc westward in a year, as observed over a period of one hundred years, and it is nowhere greater than six minutes per year in the United States. The diurnal change in the same area varies from $4'$ to $15'$ between its extremes, which occur at about 9 A.M. and 2 P.M., the needle shifting toward the east in the morning and westward in the afternoon. The annual cyclical change, and the changes due to the influence of the moon, are less than a minute of arc, and are therefore usually negligible. The effect of magnetic storms, however, is often extremely serious, amounting in some cases to declinations of several degrees, and shipwrecks have been caused by these unpredictable disturbances of the compass.

The causes of magnetic variations are very little understood, though unquestionably electrified particles shot out from the sun profoundly influence terrestrial magnetic phenomena.

SUPPLEMENTARY READING

- C. A. Culver, *Electricity and Magnetism* (Chap. 14), Macmillan, 1930.
S. P. Thompson, *Elementary Lessons in Electricity and Magnetism* (Chap. 2), Seventh Edition, Macmillan, 1926.
C. R. Underhill, *Magnets*, McGraw-Hill, 1924.

PROBLEMS

1. Calculate the force between two similar poles, of strength 24 and 18 e.m.u., when they are 12 cm apart. *Ans.* 3 dynes.
2. A bar magnet has poles of 400 e.m.u., 12 cm apart. Calculate the field strength in direction and magnitude at a point 8 cm from the center of the magnet and equidistant from its ends. *Ans.* 4.8 oersteds parallel to the magnet.

3. A bar magnet NS has poles of strength 144 e.m.u., 5 cm apart. Calculate the field strength at a point 3 cm from N and 4 cm from S . *Ans.* 18.36 oersteds.

4. A magnetized needle whose poles have a strength of 25 e.m.u. and are 3 cm apart is placed in a magnetic field of 12 oersteds at an angle of 30° with the direction of the field. What is the torque on the needle? *Ans.* 450 dyne-cm.

5. Calculate the torque required to hold a magnetized bar suspended in an east-west position in a north-south field of 0.21 oersted, when the bar has a magnetic moment of 50 (e.m.u.)-cm. *Ans.* 10.5 dyne-cm.

~ 6. The magnet of Problem 2 is placed in a uniform field of 3.6 oersteds normal to its length. What is the resultant field at the point considered, and its direction? *Ans.* 6 oersteds; $36^\circ 9'$ from the direction of the magnet.

7. Calculate the vertical field intensity of the earth's magnetism at a point where $H = 0.19$ oersted, and the inclination is 72° . *Ans.* 0.58 oersted.

CHAPTER 44

Electrostatics

572. Electrification by friction. The ancients knew that if a piece of amber or other resinous substance were rubbed with some woollen material, or still better, with cat's fur, it would attract light objects, such as bits of cloth or pith, very much as if it were a sort of magnet. A piece of dry paper rubbed over the varnished surface of a desk clings tenaciously at times. Dry glass rubbed with silk exhibits a similar effect, while the glass and silk tend to attract each other after the rubbing. All these phenomena are similar in that they result in producing a "charge" of what is known as electricity on both surfaces that have been rubbed together. The name, electricity, is derived from the Greek word *elektron*,† which means amber.

573. Two kinds of electricity. To find out more about the charges obtained by friction, let us make use of a light pith ball suspended by

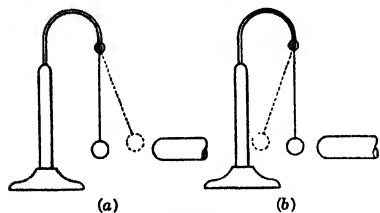


Fig. 13.

a silk thread from a convenient support, preferably of glass, as indicated. If our piece of amber, or rod of hard rubber (a common substitute), is excited by cat's fur and held near the pith ball, it will be seen to attract it, as in Fig. 13 (a), and if brought close enough, the ball will adhere. But

only for a short time. Suddenly it detaches itself and flies away under a repulsion as strong as the previous attraction, as in Fig. 13 (b). The same results (first attraction and then repulsion) will be obtained using a glass rod rubbed with silk. But now if the pith ball that is being repelled by the amber rod is brought near the glass rod instead, we shall find an attraction. Thus the ball repelled by glass will be found more strongly attracted to amber than if it were uncharged. Also, two balls treated alike repel each other, but are attracted if one ball is treated with amber and the other with glass.

Clearly we have here two sorts of electricity, and clearly also, the unlike kinds attract each other, while the like kinds repel. The

† The initial *e* should be pronounced long as in *equal*, and not as in *elbow*.

reason the amber rod first attracts and then repels is that, when the pith ball is *uncharged*, it is attracted by either kind of electricity, just as a bar of soft iron is attracted to both north and south magnetic poles. But after contact with the electrified rod, some of the electricity on it goes over to the pith ball, and then repulsion takes place between the two.

The two kinds of electricity we have thus postulated were long ago named positive and negative. Positive refers to the kind produced on glass, and negative the kind on amber, rubber, and resinous substances generally. *Vitreous* and *resinous* have also been used as names for the two kinds of charge, but this old-fashioned terminology is scarcely ever used nowadays.

574. What is electricity? It is doubtful if we shall ever know the whole answer to this, one of nature's most puzzling riddles. Still, as we learn more and more about the structure of matter we obtain a continually clearer picture of electricity, for these two entities are intimately bound up with each other. If we knew all about matter, we should know all about electricity.

Broadly speaking, matter seems to be built up out of electrically positive and negative particles. The basic negative charge is the **electron**, and the basic positive charge is most usually associated with the nucleus of the hydrogen atom, called the **proton**. This elementary positive charge also appears in certain experiments as a "positive electron," or **positron**.

We may picture any atom as made up of a central nucleus surrounded by electrons. The positively charged nucleus accounts for most of the atom's mass, and in solids the nuclei have very little mobility. In metals the electrons have considerable mobility among the atoms, and when in motion constitute an electric current. If a body contains an excess of electrons, it is negatively charged. If there is a deficit, it is positively charged.

575. Electron theory in practice. In some ways it is unfortunate that the "resinous electricity," which we now regard as meaning an *excess* of electrons, was called *negative*. However, in discussing positive and negative charges, there is no ambiguity, and ordinarily there is no need to think of them as deficits and excesses of electrons. Indeed, we save ourselves a good deal of unnecessary trouble in ignoring the electron theory when discussing charges as a whole, just as in hydraulics we talk about water and not about molecules of H_2O . That kind of refinement has its place, but need not be dragged in everywhere.

Therefore, except where the use of electrons helps make phenomena clearer, we shall use simple names and concepts in dealing with charges and currents, and the reader, if he wishes, may substitute

Positive charge = deficit of electrons.

Negative charge = excess of electrons.

Electric current = moving electrons.

The electron theory accounts in part for electrification by friction. We may assume that when two different bodies, *A* and *B*, are in contact, the free electrons (those having mobility among the atoms) from *A* tend to go over into *B* more readily than those from *B* tend to go over into *A*. Consequently *A* acquires a positive charge from loss of electrons, and *B* acquires a negative charge through acquiring an additional number. The rubbing serves only to increase this tendency by bringing more molecules of *A* in contact with those of *B* than would otherwise be possible.

576. Conductors and insulators. The production of an electric charge by friction is not limited to the substances mentioned in Article 572, but can occur between a great variety of others, provided proper precautions are taken. A ball made of brass or some other metal, if mounted on a glass rod, can be given a negative charge by cat's fur provided only the rod is grasped in the hand during the rubbing process. The metal will then act just as the amber did, but when touched by the hand, it instantly loses the power to attract. Since this is not the case with the amber or hard rubber, it shows a fundamental difference between metals and resinous or vitreous substances. The metals are *conductors* of electricity; the other substances are not. No amount of rubbing of a metallic object held in the hand will give it a charge of electricity, because both the metal and the human body are conductors, and the charge is carried off to the earth as fast as it is produced. On the other hand, when a glass rod is rubbed with a silk cloth, both rod and cloth become charged with opposite kinds of electricity because neither are conductors, and the electric charge stays on them where it was produced. Such substances are called *insulators*, and glass and silk especially have high insulating properties. No insulators however are absolutely perfect, nor are there any perfect conductors. Among the best insulators are amber, sulphur, mica, rubber, and glass, while among the best conductors are gold, platinum, copper, and silver. Moist earth is a fairly good conductor, and the human body and vegetable substances like wood conduct, though much less readily than metals. Liquids,

except liquid metals, in general are nonconductors in the ordinary sense of the word, but certain solutions carry electricity by a process known as electrolysis, which will be considered later.

577. "Grounds." It was just observed that the earth, especially when moist, is a fairly good conductor. This is an important fact, considering that most electrified bodies must in the end either give up their charges to the earth or, as we shall see, become neutralized by a like quantity of opposite electrification. In the former case, the earth may be regarded as an essentially infinite receiver of either positive or negative charges, so that electrified conductors that are in contact with the ground finally lose their charge. It is absorbed by the vast reservoir into which it flows, without appreciably affecting the electrical condition of the earth, any more than rivers raise the level of the ocean.

The walls of a house, though rather poor conductors, carry charges to the earth, but where a really good "ground" is desired, metallic conductors are necessary. Gas or water pipes may be used in this way, or a wire attached to a mass of metal buried in moist charcoal underground, such as is used at the base of lightning rods.

578. Electroscopes. Before we proceed further with the behavior of electricity at rest, it will be well to describe certain instruments which are particularly useful in observing and measuring it. The gold-leaf electroscope is perhaps the most valuable of these devices. It consists of a rod which pierces the stopper of a glass flask. It has a brass ball at its outer end, and two strips of gold leaf at the other, as shown in Fig. 14. The stopper must be of some highly insulating material such as rubber, or better still, of amber. Then when the knob receives a charge, it will not leak over to the glass, whose surface is always slightly moist and therefore very slightly conducting. This would cause a slow discharge to the ground.

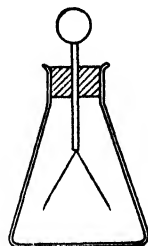


Fig. 14.

If the knob of the electroscope is rubbed with cat's fur, the charge produced on the knob is conducted all over the metallic system, including the gold leaves. Having thus acquired two small negative charges, the leaves stand out at an angle whose magnitude depends upon the amount of electrification. In this charged condition, the electroscope is a very sensitive detector of charges upon other bodies, and serves to indicate both their sign (positive or negative) and to a certain extent their magnitude.

By far the most important use of the electroscope is in measuring

the feeble conductivity which may be developed in air or other gases. Thus air can be made to conduct by means of radium rays, which create ions, or carriers of electricity, a process called ionization. The more powerful the rays the greater the resulting conductivity. A sample of radium held near a charged electroscope ionizes the air surrounding it, and the leaves collapse at a rate which indicates the amount of ionization.

579. The electrometer. In its more usual form the electrometer, represented in Fig. 15, consists of four double quadrants of brass

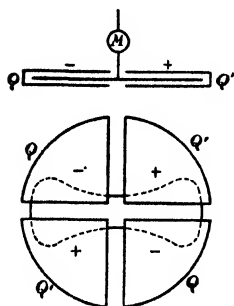


Fig. 15.

separated by a narrow air space, and mounted on insulators. Between the upper and lower portions of the quadrants a thin metal strip known as the *needle* is suspended by a conducting fiber such as silvered quartz. One pair of quadrants, QQ , diagonally opposite each other, is charged negatively, and the other, $Q'Q'$, positively. Then if the needle receives a charge through its suspension, it will be repelled from the quadrants of like sign, attracted by the others, and thus caused to rotate. A beam of light reflected from the mirror M ,

which rotates with the needle, may be concentrated as a "spot" on a scale some distance away. Very slight charges may thus be detected and measured by the amount of the deviation of the spot from its normal position. Another method, even more sensitive, is to give the needle a fairly heavy charge and connect one pair of quadrants to the charged body under investigation. The other pair is either grounded or given a like charge of opposite sign.

580. Coulomb's law. The fundamental law of electrostatics was discovered by Coulomb in 1785. It states that *the force between two very small charged spheres (considered as points) is proportional to the product of their charges and inversely as the square of the distance between them*, or

$$F \propto \frac{qq'}{r^2}, \quad (1)$$

where q and q' are the charges, and r is the distance between them. When q and q' are both positive or both negative, F is positive, and the charges repel each other. When q and q' are of opposite sign, F is negative, and the charges attract each other. Coulomb established this law by means of an apparatus such as that shown in Fig. 16. It consisted of a torsion balance having a small metal sphere at the

end of a light nonconducting bar which was acted on by a fixed metal sphere of the same size. When both spheres were in contact and charged, the force of repulsion twisted the arm through a definite angle α . The angle β , through which the "head" must be turned to diminish α by a given amount, served to compare the forces acting at different distances between the spheres. The effect of varying q and q' was studied by touching an uncharged sphere, of the same size as those in the balance, to either the movable or fixed sphere. This resulted in halving its charge, and the force was then found to be half as great as before. A second application reduced it to one fourth, and so on.

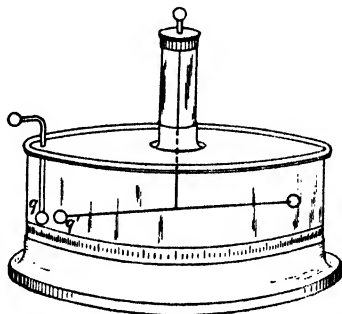


Fig. 16.

581. Unit of charge. From Coulomb's law we may define the absolute unit of an electrostatic charge in the so-called electrostatic system of units (or e.s.u.) so as to write relation (1) above as an equation. Let the distance between charges be one centimeter, the force one dyne, and the medium be a vacuum; then if both charges are equal, they must be of unit value. We may therefore define unit q as follows: *The unit charge (denoted by esu)† is one which, in a vacuum, repels an equal and like charge at a centimeter's distance with the force of one dyne.*

Since unit q has now been defined so as to make F numerically equal to qq'/r^2 , we may state Coulomb's law as

$$F = \frac{qq'}{r^2}. \quad (1)$$

This relation gives us a means of obtaining the dimensions of q , which must be the same as those of a magnetic pole m , since Coulomb's law has the same form for both. Therefore

$$[q] = [M^{\frac{1}{2}}L^{\frac{3}{2}}T^{-1}].$$

582. Induced charges. According to modern views of matter, all uncharged bodies contain within themselves the elements of both

† The abbreviations emu and esu mean *electromagnetic and electrostatic units of quantity*, respectively. The abbreviations e.m.u. and e.s.u. mean *electromagnetic and electrostatic units*, and may be applied to any of the various absolute units in these sections, such as those of current, resistance, and so forth.

positive electricity (atomic nuclei) and negative electricity (electrons) in equal amounts. If an uncharged conductor is brought near a positively charged body, for example, the negative electricity in the conductor flows to the end nearest the positive charge, and leaves the far end positive. If the body is charged negatively, the negative electricity in the conductor is repelled and leaves the near end positive. This process is called **induction**.

Induced charges may be readily exhibited with an electroscope. If a negatively charged rod is brought near the knob, the leaves diverge with the repelled negative charge, but collapse again when the rod is removed. If, however, the knob is grounded by placing a finger on it while it is under the influence of the rod, the leaves collapse, owing to the flow of the repelled negative charge to the earth. The knob, having lost negative electricity (electrons), is positively charged, and remains so under the influence of the negative electricity on the rod. Then if first the finger and next the rod are withdrawn, the leaves diverge once more. The charge is positive and is no longer confined to the knob, but spreads itself all over the metallic system. This is much the best method of charging an electroscope, although it can also be done by simple contact with a charged body.

The charged electroscope is used to identify charges by their inductive action. If it is charged positively (electron deficit), a positive charge held near the knob causes the leaves to diverge still more, since it attracts negative electricity into the knob and leaves a still greater deficit of electrons on the leaves. A negative charge makes the leaves collapse by repelling negative electricity from the knob into the leaves. This tends to neutralize the positive charge on them by making up the deficit of electrons. A negatively charged electroscope behaves in an exactly opposite manner.

583. All free charges superficial. Benjamin Franklin, in 1780, by an indirect experiment, and Faraday, in 1837, by a direct one, proved that electric charges reside wholly on the surface of conducting bodies. Faraday's experiment is classic, and consisted in having himself shut up in a metal box which was then charged. He was unable to detect any trace of electricity in the interior, although sparks were drawn from the box by observers outside.

This fact can be demonstrated in a variety of ways, but could have been predicted from the known repulsion between like charges. This would naturally cause their constituent elements (excess electrons, for instance) to get as far apart as possible, such as occurs when these elements are on the outer surfaces of charged bodies.

584. Faraday's ice-pail experiment. Another classic experiment was performed by Faraday, with a view to determining the exact laws of induction. He lowered a charged sphere by a nonconducting string into a metal pail that was connected by a wire to the knob of an electroscope. A metal cover should be attached to the string, as shown in Fig. 17, if the experiment is to be conclusive. As the ball with, let us say, a positive charge, is lowered into the "pail" the leaves begin to diverge with a loss of negative electricity, and consequent positive charge. They reach an observed maximum angle of separation when the cover closes the opening. If the ball is withdrawn, the leaves collapse.

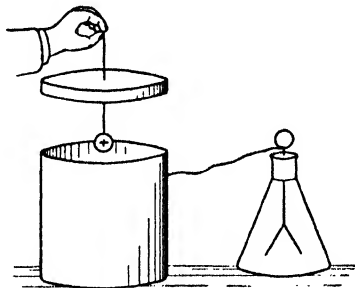


Fig. 17.

Repeating the process, we may cause the ball to strike against the side of the pail when it is closed, but this does not alter the divergence of the leaves. However, upon withdrawal, the ball is found to be discharged, and the leaves remain deflected by the same amount as before.

These phenomena show that the charge attracted to the inside of the pail must have been equal in strength to that on the ball, because contact resulted in complete neutralization, leaving the repelled charge on the electroscope constant, as suggested by the number of plus and minus signs in Fig. 18. The fact that the leaves collapse when the ball is removed without contact shows that the two induced charges must also have been equal. So we are now justified in concluding that when a charged body is completely surrounded by a metal shield, it induces two opposite charges in that shield each exactly equal to its own. The attracted charge is on

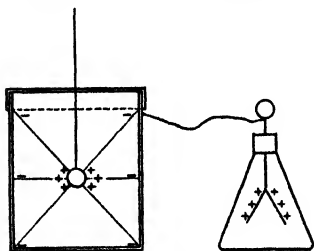


Fig. 18.

the inner surface only so long as it is *bound* there by the charge on the ball, and it is therefore not subject to the law that *free* charges reside on the outer surfaces of bodies.

A further confirmation of these principles can be obtained by grounding the pail while the ball is inside it, and then removing the

ball. The leaves collapse upon grounding, owing to the escape of the repelled charge, and then diverge to the same extent as before when the ball is withdrawn, showing the equality of the two induced charges.

585. Electrostatic fields. Field strength, or intensity, is defined as *the force in dynes acting upon a unit positive charge at a specified point*. This quantity, denoted by E , may be calculated by dividing the force exerted by the field upon a given charge q' by the amount of that charge, or $E = F/q'$. Then the force F in a field E is given by

$$F = Eq'. \quad (1)$$

Using Coulomb's law, we find the field strength at a point p , distant r cm from a charge of q esu, by

$$E = \frac{F}{q'} = \frac{qq'}{r^2q'} = \frac{q}{r^2}. \quad (2)$$

Actually the charge q' disturbs the field in which it is placed, but the result, force per unit charge, is really an ideal case independent of any actual charge at p .

To calculate the field strength at p due to any number of charges, positive or negative, we proceed exactly as in magnetism, and obtain a vector sum, as shown in Fig. 4, Article 559, though now $+q$ replaces any north-seeking pole, and $-q$, any south-seeking pole. Then the field strength at p due to a positive charge $+q$ is given by $E_1 = q_1/r_1^2$, and a negative charge gives $E_2 = -q_2/r_2^2$. Here the two charges need not be equal, as in the case of a magnet, but the *resultant*, as before, is the vector sum of the intensities, or $\overline{E_1 + E_2}$.

The dimensions of E are the same as those of H . This is evident because m and q have the same dimensions, and the field intensities are defined in the same way. The dimensions of E then are

$$[E] = [M^{\frac{1}{2}}L^{-\frac{1}{2}}T^{-1}].$$

Unit field strength may be defined either by means of equation (2) as esu/cm², or by means of the definition of E , as dyne/esu. The choice depends upon whether the charge in question does or does not create the field.

586. Electrostatic lines of force. As in magnetism, an electrostatic field may be mapped out with lines of force. These lines are defined in exactly the same way, and it is assumed that 4π lines radiate from a unit charge. It follows, as proved for magnetic fields (Article 560), that in a field of intensity E , the number of lines of force cutting across a square centimeter normal to their direction is numerically equal to E .

587. Electrostatic potential. Like all potentials, the absolute potential V at a point in an electrostatic field is a space property of that point, and is measured in terms of the work required to carry a unit quantity from a region of zero potential to the point in question. In this case the unit quantity is the unit positive charge, and the zero region is infinitely far from any charge. Or we may consider the potential due to a particular charge, when the zero region is simply so far away from that charge that its field there is negligible.

In most problems, however, we are not interested in absolute potential, but in the potential difference, $V_2 - V_1$, between two points. Then this potential difference is measured by the work required to move a unit charge from one point to another against the force acting upon it in an electrostatic field. In a uniform field in which E is constant, we may write

$$V_2 - V_1 = \frac{\text{Work}}{q'} = \frac{Fs}{q'} = Es,$$

where s is the distance between the points.

If the field is due to an isolated charge, it is not uniform. Then we calculate the potential difference as follows: In Fig. 19 let p_1 and p_2 be

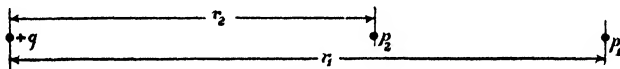


Fig. 19.

two points whose distances from a positive charge q are r_1 and r_2 . The field intensities at p_1 and p_2 are $E_1 = q/r_1^2$, and $E_2 = q/r_2^2$, respectively, as shown in equation 2, Article 585. The average intensity is the geometrical instead of the arithmetical mean of these two values, because they depend upon the inverse square of the distance, rather than directly upon its first power. Therefore $E_{av} = \sqrt{E_1 E_2} = q/r_1 r_2$. The work per unit charge, W/q' , found by moving a positive charge q' from p_1 to p_2 against the field, gives the potential difference between these points. Therefore, as V_2 is greater than V_1 ,

$$\begin{aligned} W/q' = V_2 - V_1 &= E_{av}(r_1 - r_2) \\ &= \frac{q}{r_1 r_2}(r_1 - r_2) \\ &= q\left(\frac{1}{r_2} - \frac{1}{r_1}\right). \end{aligned} \quad (1)$$

This is the difference of electrostatic potential between these points. If p_1 is infinitely distant, the field due to q is zero, and its potential V_1 ,

as influenced by q , is zero also. Then, setting $r_1 = \infty$, we obtain the work required to bring a unit charge from a point of zero potential to one where the potential is V_2 . This is obviously equal to q/r_2 , or in general, the potential due to a charge q is given by

$$V = q/r. \quad (2)$$

Unlike field intensity, electrostatic potential, or in fact any potential, is a scalar quantity. This is because potential is based on work done in opposing a force, and work has no direction.

The fact that potential is not a vector makes calculation of potential due to several charges extremely simple. We need take only the algebraic sum of the various potentials at a given point. Thus in

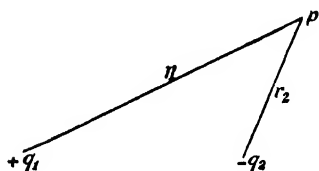


Fig. 20.

Fig. 20, the potential at p due to the charge q_1 is $+q_1/r_1$, and that due to q is $-q_2/r_2$. The negative sign indicates that work would be *done by* a unit positive charge in bringing it from a region of zero potential to p when only the field due to q is considered. Then

$$V = q_1/r_1 - q_2/r_2.$$

The potential energy of two charges, q and q' , at a distance r from each other, is the work done in bringing one of them, say q' , from a region of zero potential to that point. This is given by

$$W = qq'/r, \quad (3)$$

because q/r is the work per *unit* charge, but when the charge is q' , the work is q' times as large.

The dimensions of electrostatic potential are those of work divided by a charge; hence

$$[V_E] = \left[\frac{ML^2T^{-2}}{M^{1/2}L^{1/2}T^{-1}} \right] = [M^{1/2}L^{3/2}T^{-1}].$$

The unit of V_E may be obtained from the ratio W/q , in which case it is erg/esu, or it may be obtained from the potential due to a charge q , that is, q/r . In this case it is esu/cm.

588. Potential gradient. An important relation connecting potential and field intensity is derived as follows: If a force F acts through a vanishingly small distance ds , the work done is given by $dW = Fds$. But in electrostatics, since $V_2 - V_1 = W/q$, it follows that $-dV = dW/q$; therefore

$$-dV = Fds/q. \quad (1)$$

Since $E = F/q$ by definition, the force F on the charge q equals Eq . Substituting this value for F , we obtain

$$-dV = Eds,$$

and

$$E = -\frac{dV}{ds}. \quad (2)$$

Equation (2) may be expressed in words by the statement that *the intensity of an electrostatic field at any point is equal to the space rate of change of potential at that point.*

The quantity dV/ds is often called the **potential gradient** of the field, and is a valuable aspect of the field intensity E . Thus E is seen to measure the steepness of the potential slope. It is similar to the "grade" in surveying, where it is expressed as a rise or fall of so many feet in a hundred. In electrostatics it serves to indicate the force on a unit charge as measured by the closeness of the lines of equal potential or electrostatic level, corresponding to contour lines on a surveyor's map.

589. The potential of a charged sphere. Suppose a charge q to be distributed over a sphere of radius r . As discovered by Faraday, a charge on any closed conducting surface distributes itself in such a way that the electrostatic force within the enclosure is zero. In the case of a sphere, the distribution is uniform. With no interior field, no work would be done in moving a charge from the center to the surface. The surface is therefore at the same potential as the center, which is obviously equal to q/r , in accordance with equation (2) of Article 587. In order, then, *to find the potential of an electrified sphere, divide the charge by the radius.* At any point outside the sphere the potential is the same as if its charge were concentrated at its center. This is obvious, in case the point is actually on the outer surface, for then it has the same potential as the sphere itself, or q/r . Then if the sphere should shrink upon its center, retaining the same charge, the potential at the original distance r would still be q/r . We may start with any radius in this imaginary process, and both at the start and finish the potential is q/r . Therefore q/r is the value of the potential at any point on or outside of the sphere at a distance r from its center. If the point lies inside the sphere, the potential is the same as if it were on the surface, as has been indicated.

590. Problems involving potential. In such problems it is necessary to calculate the potential at a point in space or on a body, taking account of all the charges which influence it, including its own, if it

has one. As an illustration, suppose a sphere of radius 4 cm, charged with 12 esu, and two charges of $+100$ and -80 units situated at

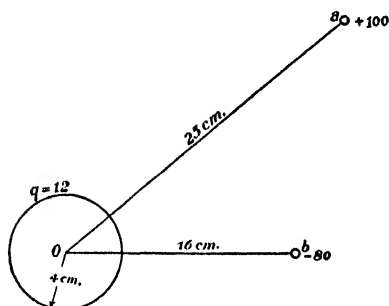


Fig. 21.

25 cm and 16 cm from its center in any direction, as shown in Fig. 21. Then the potential at O (the center of the sphere), due to a , is $+4$, and to b , -5 units. If the sphere were uncharged, the potential at O would be -1 , but it has a potential of $+3$ due to its own charge, so its net potential due to all influences is $+2$.

SUPPLEMENTARY READING

S. P. Thompson, *Elementary Lessons in Electricity and Magnetism* (pp. 1-44), Seventh Edition, Macmillan, 1926.

A. Zeleny, *Elements of Electricity* (Chapters 6, 7), McGraw-Hill, 1930.

PROBLEMS

1. Two pith balls in contact, each weighing a decigram, and suspended in air by fine fibers 30 cm long, are given equal like charges. They are repelled to a distance of 18 cm from each other. What is the force of repulsion? What is the charge on each ball? *Ans.* 30.8 dynes; 99.9 esu.

2. Two small metal spheres of the same radius have charges of $+15$ and -9 units. What is the force of attraction between them when their centers are 6 cm apart in air? What is the force of repulsion when they have been placed in contact and then restored to their original positions? *Ans.* 3.75 dynes; 0.25 dyne.

3. Two small bodies having charges of $+1000$ and -216 esu are 16 cm apart in air. Calculate the field strength and its direction at a point 12 cm from the negative charge and 20 cm from the positive one. *Ans.* 2 dynes/esu and parallel to the line connecting the charges.

4. Calculate the potential at the point specified in Problem 3. *Ans.* $+32$ erg/esu.

5. What is the potential of a sphere of 4 cm radius charged with $+18$ esu and placed 18 cm from a charge of -45 esu? *Ans.* $+2$ erg/esu.

6. What is the potential of the sphere in Problem 5 when a second negative charge of 12 units is placed 6 cm away? What if the new charge were positive? *Ans.* Zero; $+4$ erg/esu.

7. How much work is done in moving a charge of 8 units a distance of 30 cm against a field which increases uniformly from 12 to 50 lines/cm² over the distance, and what is the average potential gradient? *Ans.* 7440 ergs; 31 dyne/esu.

CHAPTER 45

Electrostatics (*Continued*)

591. Mapping electrostatic fields. The region around a charged body, wherever its influence is felt, is known as its **field**. This may be graphically indicated, as in magnetism, by lines of force. Lines of force are conceived as starting upon the surface of a positively charged body and terminating upon a negatively charged body. Their course indicates the direction of the force upon a positive charge placed at that point, so they are directed away from the positive charge where they originate.

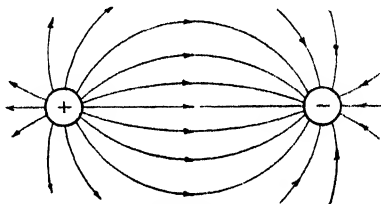


Fig. 22.

A field mapped out in such lines shows graphically both the direction and intensity of the field, as indicated in Fig. 22. Where the lines are crowded together, the field is strongest; where they are most separated, it is weakest.

Since two unlike charges attract each other, these lines may be thought of as stretched elastic bands tending to shorten indefinitely and so pull the two charges together. But, unlike any known elastic

substance, the pull gets stronger as they grow shorter. This apparent paradox is shown graphically by their tendency to separate as much as possible as they become stretched, which may be regarded as equivalent to mutual repulsion. Thus when two charges are not far apart, the lines of force connecting them are nearly straight, but as they



Fig. 23.

are separated, the lines bow out more and more, as indicated in Fig. 23. The more they are curved, the less direct is the pull, and the force urging the charges together is weaker.

592. Surface density. The quantity of electricity on each square centimeter of the surface of a charged body is known as **surface density**, and is measured by the charge divided by the area. If the distribution is uniform, the surface density is expressed as $\sigma = q/A$; otherwise this gives only average density, and we should have to write $\sigma = dq/dA$ for the density at a specific point in an area where the distribution varies from point to point.

593. Conductors and insulators in electrostatic fields. Suppose an insulated conducting body (a metallic sphere for example) is placed in the field caused by a charged surface. If this field is then examined by means of an electroscope or any suitable device, it will be found that the lines of force converge toward the conductor and are more concentrated both on the side facing the charged surface and on the opposite face. Further, the two faces exhibit induced charges of opposite sign, as indicated in Fig. 24. No lines of force are shown

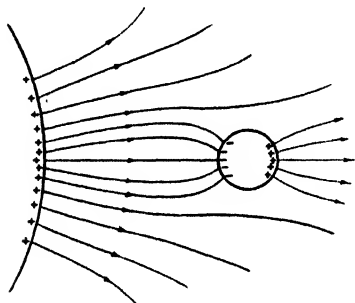


Fig. 24.

going through the conductor, because there is no field within a charged conducting surface, as has already been stated.

The induced negative charge shown in Fig. 24 is possible because within a conductor are many free electrons which are drawn toward the positively charged body. The induced positive charge is due to a deficit of electrons drawn away from positive nuclei.

In the case of nonconductors, substances such as amber, glass, hard rubber, and so forth, are better media for an electrostatic field than air. These are called **dielectrics**. When placed in a field, the lines of force tend to become concentrated in the nonconductor, as shown in Fig. 25,

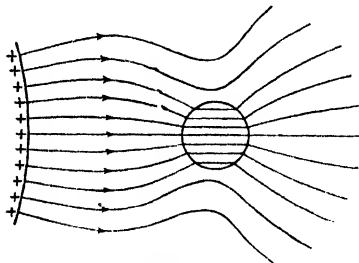


Fig. 25.

because it offers them a better path than the surrounding air. In the case of conductors, the charges are induced on the surface, and consequently the lines of force end there; but since no charges are induced on the surface of a nonconductor, the lines of force pass

straight through. The reason no charges are induced in such substances is that there are no free electrons to be drawn toward a positively charged body or repelled from a negative charge.

It will be seen that with both conductors and nonconductors the force between the two bodies is an attraction, because, regarding the lines of force as stretched elastic bands, they must exert a greater resultant force on the side toward the charged body, where they are nearly parallel, than on the other, where their divergence is more marked. The result of this attraction tends to cause motion into a field of greater intensity, with still greater concentration of the lines of force. In a uniform field having strictly parallel lines of force, neither attraction nor repulsion can take place, for there is no difference in the concentration on the two faces of the body we are considering. However, even in this case, an elongated dielectric body in air, like a paramagnetic body in a magnetic field, would set itself parallel to the lines of force.

There is no analogue of diamagnetism in electrostatics. That is, there are no substances which are poorer media for lines of force than a vacuum. Therefore, in a vacuum, elongated specimens of dielectrics would always set themselves parallel to the lines of force.

594. Equipotential surfaces. In addition to the lines of force serving to give graphic representation of the *field of force* around an electric charge, we may also plot other lines, the traces of surfaces, along any one of which the *potential* is constant. These are exactly analogous to the contour lines in a geodetic map, indicating in the one case constant electrostatic potential, and in the other, constant gravitational potential, or level. Such surfaces intersect lines of force at right angles, and with the latter are said to constitute an **orthogonal system**. This can be proved by the "reductio ad absurdum" method as follows: Suppose a line of force cd

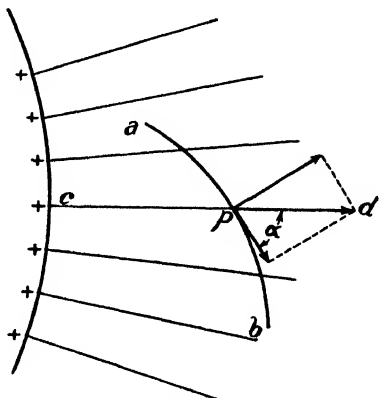


Fig. 26.

cuts an equipotential surface ab at an angle α less than 90° . We may then decompose an element of the line of force into two components, one normal and one tangent to the equipotential surface, as indicated in Fig. 26. The component tangent to the equipotential surface

coincides with it at p , which indicates that work must be done at that point in moving a charge along ab . But this is contrary to the hypothesis, since by definition an equipotential surface of any sort is so drawn that no work is done in moving the unit concerned (in this case a charge) along it. Therefore α must be 90° , because only then can cd have no component coinciding with ab .

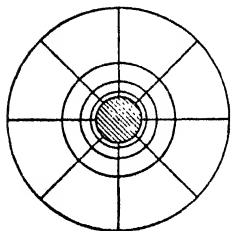


Fig. 27.

595. Orthogonal systems. The combination of intersecting lines of force and equipotential lines representing surfaces, as indicated in Fig. 27, gives a complete survey of the conditions of an electrostatic field due to one or more charges. For a charged sphere they form a symmetrical pattern in which the equi-

potential lines are more closely crowded as they approach the surface, just as are the lines of force. If the sphere has a radius of 2 cm, for instance, and a charge of 12 units, its potential is 6. The radius of the line indicating a level of 5 is obtained from $V = q/r$, or

At the level where $V = 5$, $r = \frac{1}{5}^2 = 2.4$ cm.

At the level where $V = 4$, $r = \frac{1}{4}^2 = 3$ cm.

At the level where $V = 3$, $r = \frac{1}{3}^2 = 4$ cm.

At the level where $V = 2$, $r = \frac{1}{2}^2 = 6$ cm.

At the level where $V = 1$, $r = \frac{1}{1}^2 = 12$ cm.

At the level where $V = 0$, $r = \frac{1}{0}^2 = \infty$.

If the above figures were from the survey of a mountain, it would appear as shown in Fig. 28. It has a flat summit, say 6 miles above the sea, and falls off very precipitously at first, but then more and more gradually, until at an infinite distance from the peak it would reach sea level.

596. Electrostatic "contour" maps. The lines of force in systems such as we

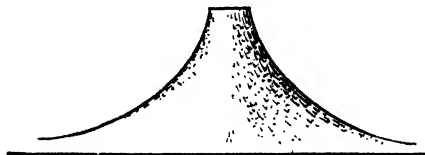


Fig. 28.

have been discussing, whether electrostatic or gravitational, indicate the route an object would take if left to itself under the action of the electrostatic or gravitational field, provided it moved so slowly as not to acquire any appreciable momentum. A positive charge in one case, or a freely sliding mass in the other, would move from high to low potential along such lines until it reached the plane of zero potential.

If we consider the special case of two equal like charges, we obtain a diagram in which the lines of force are curved and the equipotential lines are no longer circles, as in Fig. 29. There is one limiting equi-

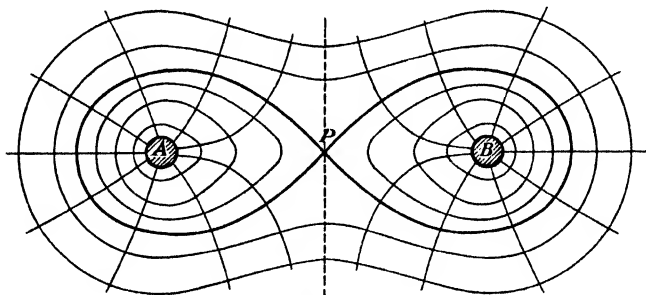


Fig. 29.

potential line forming a figure eight. Outside of this line the contour lines embrace both peaks, but inside of it they surround only one. Intersecting the center of the "8" is a line which does not originate in either charge, but forms a sort of limit for both sets of lines of force. The point P of the intersection is one of unstable equilibrium for any charge placed there.

If the equal charges are unlike (Fig. 30), the reverse is the case. Here one is analogous to a hole, and tends to steepen the side of the

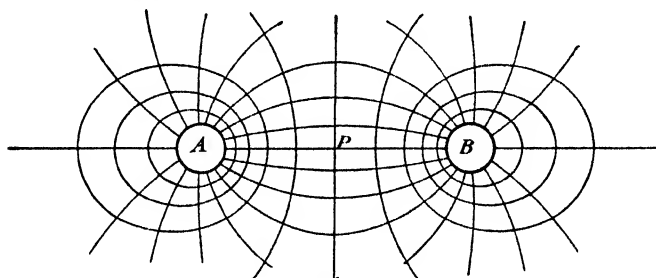


Fig. 30.

"peak" nearest it. Also one of the lines of force is straight, showing the shortest path a charge would take in passing from A to B , while there is a limiting contour line which is also straight and at zero potential.

Contours such as we have been discussing may readily be produced by stretching a sheet of rubber over a horizontal frame and then forcing a blunt rod either up or down at any convenient point. The

influence of such elevations or depressions on the level (potential) of the surrounding surface is most instructive, especially if two or more are used simultaneously.

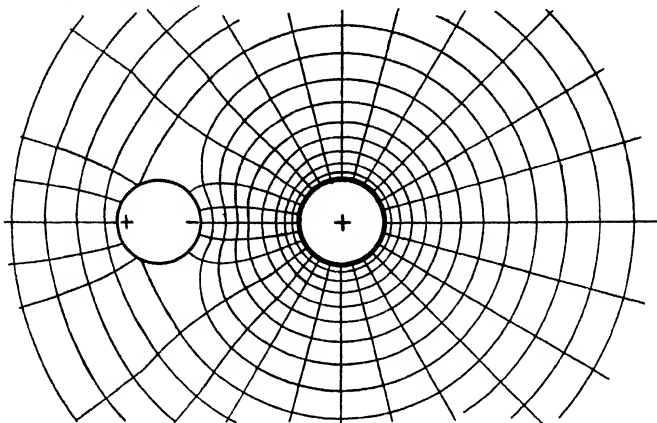


Fig. 31.

The orthogonal system associated with induced charges is shown in Fig. 31. Here a single equipotential surface is associated with the conductor. It includes the contour of the conductor, and thus forms a sort of bulge, any part of which is at the same potential as the rest of the surface. Also, the crowding together of these surfaces between the bodies indicates a concentration of electrostatic force there, while outside the conductor the gradient is less steep.

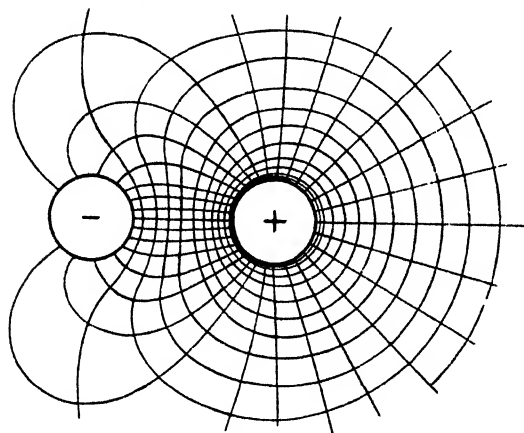


Fig. 32.

If the conductor is grounded while under the influence of the charge, the repelled electricity goes to earth, leaving only

the attracted charge, and the diagram is altered by the disappearance of the lines emanating from the repelled charge, and by a redistribu-

tion of those ending in the attracted charge. This is due to a greater spread of the attracted charge over the sphere, as shown in Fig. 32. The equipotential surfaces are now more crowded together near the conductor. This means that the conductor is being urged toward the charge with a greater force than before, as we should expect.

597. Distribution of charges on conductors. This problem is generally insoluble, but in the case of a few simple solid figures, like an ellipsoid of revolution, we may calculate how the density varies over the surface. The calculation shows that the density tends to increase with the curvature, a result amply verified by experiment. The reason for this may be made clear by the following experiment: Let two spheres, of radii r_1 and r_2 , connected by a long fine wire, have charges q_1 and q_2 units respectively. Since they are in contact through the wire, they have the same potential. Therefore

$$V = \frac{q_1}{r_1} = \frac{q_2}{r_2}$$

$$\text{and} \qquad \frac{q_1}{q_2} = \frac{r_1}{r_2}. \qquad (1)$$

Now disconnect the wire, keeping the charged spheres so far apart that each may be regarded as uninfluenced by the other. As the charges are then distributed uniformly over the two surfaces, the densities are

$$\sigma_1 = \frac{q_1}{4\pi r_1^2} \quad \text{and} \quad \sigma_2 = \frac{q_2}{4\pi r_2^2}.$$

The ratio of these densities is

$$\frac{\sigma_1}{\sigma_2} = \frac{q_1 r_2^2}{q_2 r_1^2}. \qquad (2)$$

Substituting the value of q_1/q_2 from (1) in (2), we obtain

$$\frac{\sigma_1}{\sigma_2} = \frac{r_2}{r_1}. \qquad (3)$$

Therefore the densities are inversely proportional to the radii of the spheres, which means that they are directly proportional to the curvatures.

This conclusion is of great practical importance because it tells us that charged points and fine wires are seats of very intense electrification, and so tend to discharge into the air. The discharge originates

in free ions (electrically charged particles) always present in the atmosphere. These are driven with great violence by the field near the point, so that they produce new ions by colliding with neutral atoms or molecules. The air near the point is thus filled with carriers of electricity which serve to unload the charge almost as if it had been grounded.

598. Capacitance. This quantity is a measure of the ability of a conductor to retain a charge, and is defined as the ratio of the quantity of electricity to its potential, or

$$C = \frac{q}{V}. \quad (1)$$

It is exactly analogous to the section of a cylindrical pail of indefinite height, while the level to which a given volume of liquid poured into it will rise is the equivalent of the potential. Obviously, the lower the level attained by a given quantity of liquid, the greater the "capacity" of the pail. We may also regard its capacity as measured by the amount of liquid required to fill it to a certain height. In electrostatics a large capacitance means a low potential, when a given charge is communicated to a conducting body.

599. Capacitance of a charged sphere. In this particular case it is easy to calculate the capacitance, because if the sphere is charged with q units of electricity, its potential, as we have seen, is q/r ; but as $C = q/V$,

$$C = \frac{q}{q/r} = r. \quad (1)$$

Thus the capacitance of a sphere, in air, is numerically equal to its radius. The electrostatic unit of capacitance has therefore the dimensions of a length and is equal to one centimeter.

600. Condensers. The above calculation assumes the sphere to be far removed from the influence of any other charge. But if such a charge, induced or otherwise, exists in its neighborhood, its capacitance is profoundly altered. When this results in an increase of capacitance, the system is called a **condenser**. As a simple case, imagine two concentric shells of radii r_1 and r_2 , as shown in Fig. 33, and separated by air. Then if $+q$ units are applied to the inner shell, its potential will be given by $V_1 = +q/r_1$, and if the outer shell is grounded, it becomes charged with $-q$ units and will have a potential $V_2 = -q/r_2$. But the total potential as measured from the

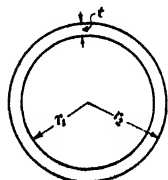


Fig. 33.

center (where both charges may be regarded as concentrated) is

$$\begin{aligned} V_1 + V_2 &= \frac{q}{r_1} - \frac{q}{r_2} \\ &= q \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \\ &= q \left(\frac{r_2 - r_1}{r_1 r_2} \right). \end{aligned}$$

Therefore, the capacitance is given by

$$C = \frac{q}{V_1 + V_2} = \frac{r_1 r_2}{r_2 - r_1}. \quad (1)$$

In case the distance t between the spheres is very small compared to their radii, we may set $r_1 r_2 = r^2$, where r is the radius of either sphere, and $r_2 - r_1 = t$. Then (1) reduces to the simple form

$$C = \frac{r^2}{t}. \quad (2)$$

Multiplying both numerator and denominator of the second member by 4π , we obtain

$$C = \frac{4\pi r^2}{4\pi t}.$$

But $4\pi r^2$ is the area A of the sphere's surfaces; therefore

$$C = \frac{A}{4\pi t}. \quad (3)$$

As the charge is uniformly distributed over the sphere, we may let A stand for any portion of the total surface, such as the cap shown in Fig. 34. Then the capacitance of such a portion is still obtained from the formula above. Further, if we imagine r to increase indefinitely, the area considered becomes a plane surface at the limit where $r = \infty$. Hence a condenser made of two plane parallel plates still has a capacitance given by equation (3), where A is the area of the plates and t is the thickness of the air space between them.



Fig. 34.

601. The Leyden jar. This is the earliest type of condenser, and derives its name from the city of Leyden, Holland, where it was invented in 1745. It is a cylindrical bottle coated inside and out with tinfoil usually to about two thirds of its height. A brass rod with a

knob at one end and a short chain at the other passes through an insulating stopper, as shown in section in Fig. 35. The chain serves to connect the inner coating to the knob, which is used in charging and discharging the jar. A coat of shellac covers the uncoated portion of the bottle to prevent the creeping of the charge over the generally moist surface of the glass. To charge the jar the outer coating is grounded, either by the hand or by any convenient contact with the earth, and the knob is connected to the terminal of some machine for producing electric charges. Its capacitance is calculated by the same formula as for spherical or plane plate condensers.



Fig. 35.

602. Specific inductive capacity. It was discovered by Faraday that if the space between the spherical shells, described in Article 600, is filled with some dielectric such as paraffin, the capacitance of the condenser is increased. This change is measured by comparing two such condensers made alike, but with the space in one of them filled with air, and in the other with the substance to be investigated. The air condenser is charged to a potential V_1 , and thus acquires a charge $C_a V_1$. The two condensers are then connected, bringing both to a new potential V_2 . The air condenser now has a charge $C_a V_2$, and has lost $C_a V_1 - C_a V_2$ esu. This equals the charge $C_x V_2$ given to the other condenser. That is,

$$C_x V_2 = C_a V_1 - C_a V_2,$$

or

$$\frac{C_x}{C_a} = \frac{V_1 - V_2}{V_2}. \quad (1)$$

The ratio of the capacitances is thus found in terms of the observed values of the potentials.

The property of the medium which increases the capacitance of one of the condensers is called its **specific inductive capacity**, or **dielectric constant**, and may be denoted by K . As the capacitance depends upon K , we may express this fact by the simple relation

$$\frac{K_x}{K_a} = \frac{C_x}{C_a}. \quad (2)$$

But the dielectric constant of air (strictly speaking, a vacuum) is arbitrarily taken as unity; therefore (2) becomes

$$K_x = \frac{C_x}{C_a}. \quad (3)$$

Thus K may not only be measured but also defined as the ratio of the capacitances of two condensers in one of which the dielectric is air, or more strictly, a vacuum. These condensers need not be like Faraday's concentric spheres, for the same relation holds true for any two similar condensers, such as those made of parallel plates separated by a dielectric.

The results of experiment show quite a range in the values of dielectric constants, all the way from practically unity for air to 81 for distilled water. The following table gives values of K for some of the more important substances, and shows why the best condensers have mica or glass as their dielectric. They have the highest dielectric constants of any of the common solid substances.

In addition to its effect on capacitance, the constant K has an important bearing on electrostatic forces and potentials. Experiment shows that in a medium whose dielectric constant is K , Coulomb's law becomes

$$F = \frac{qq'}{Kr^2}, \quad (4)$$

whence it follows that the field strength due to a charge q is given by

$$E = \frac{q}{Kr^2}, \quad (5)$$

while the potential due to q becomes

$$V = \frac{q}{Kr}. \quad (6)$$

Thus all three quantities are diminished when K is greater than unity. The decrease in V accounts for the increase in the capacitance of a condenser, for since $C = q/V$, a smaller V means a larger C , and equation (3) of Article 600 becomes

$$C = \frac{KA}{4\pi t}. \quad (7)$$

Substance	Dielectric Constant (K)
Crown Glass.....	5 to 7
Flint Glass.....	7 to 10
Rubber.....	2.1 to 2.3
Mica.....	5.7 to 7
Paraffin Wax...	2 to 2.3
Paraffin Oil.....	4.6 to 4.8
Sulphur.....	3.6 to 4.3
Ethyl Alcohol...	26.8 at 14.7° C
Water.....	81.
Air.....	1.000586 at 0° C

603. Residual discharge. After a condenser having a solid dielectric has been apparently completely discharged, it is possible to obtain several more sparks from it, provided a reasonable length of time has elapsed between each discharge. This suggests a gradual yielding of the electrostatic strain to which the dielectric was subjected when it became the seat of a field of force. Evidently the conducting surfaces are not the true seat of the energy represented by the charge, but are only the limiting planes of a field which exists in the intervening material. This strain is largely released when the first discharge takes place, but a residuum remains, just as a bent bar of imperfectly elastic material does not straighten out all at once, but may gradually do so if it has not been bent too far.

604. Energy of an electrostatic charge. All charges, whether on an isolated body or a condenser, represent energy. Since the electricity is at rest, this must be potential energy that is capable of being transformed into the kinetic energy of the discharge. In general, potential energy represents work previously done on a system against an opposing force. As we saw in Article 587, potential difference ($V_2 - V_1$) is equal to W/q . Then the work done in moving any charge q in a field of force from one potential level to another level is given by $W = q(V_2 - V_1)$. If the first level is zero, the final energy of the charge q is Vq , where V is the potential to which q was raised. But if a conductor or condenser is uncharged, its potential level is zero and the work done in bringing up an initial small quantity is negligible. As the charging process continues, the potential rises and the work of bringing up additional charges increases exactly as the work of building a tower, or filling a stand pipe with water pumped in from below, as was pointed out in Article 65. Therefore, in computing the total work done, represented by the potential energy of the charge, we must take the average potential, which is half of the final value. Then

$$W = Vq/2, \quad (1)$$

where V is the final potential and q is the total charge. If the capacity of the charged body or condenser is C , then, since $C = q/V$, we may eliminate either V or q , and obtain

$$W = V^2C/2 \quad (2)$$

and

$$W = q^2/2C. \quad (3)$$

All three expressions are useful in calculating the energy of an electrostatic charge, according to which two of the quantities V , q , and C are given.

605. The electrophorus. The essential idea involved in most machines for producing electrostatic charges can be best understood by a study of a device known as the **electrophorus**, shown in Fig. 36. It consists of a hard rubber disc mounted on a metal plate which is in contact with the ground. A second metal disc of the same size is fitted with a hard rubber handle by which it can be raised or lowered over the rubber plate. The latter is then "excited" by rubbing it with cat's fur, and acquires a negative charge on its upper surface, while the lower layers act like the dielectric of a condenser. This results in an induced positive charge on the upper surface

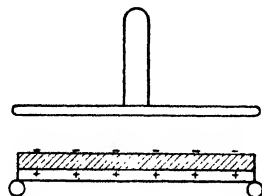


Fig. 36.

of the supporting metal plate. The movable plate is then lowered into contact with the rubber disc, and thus acquires induced positive and negative charges on its lower and upper surfaces respectively. This does not appreciably remove the negative electricity from the rubber, because in reality the contact is made only at a few points, and even if these are discharged, there is no flow over this nonconducting surface to carry off the entire charge. The situation is now as indicated in Fig. 37. The next step is to ground the movable plate, which is most effectively done by connecting it to the lower positive charge of the system. This leaves the two middle charges in close contact and at zero potential with respect to the earth.

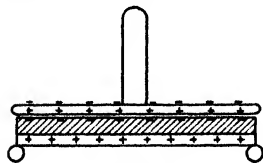


Fig. 37.

If the upper plate is now lifted by the handle, its potential is raised, while its positive charge is progressively removed from the influence of the negative field. When a foot or so away, the charge on the metal disc has a potential sufficiently above the earth to discharge in a spark an inch or more long to the hand held near it. Meanwhile the rubber disc has fallen in potential to a point as far below zero as the plate was above it, and causes a positive charge to flow from the earth into the lower plate. The whole process may now be repeated, and a succession of sparks may thus be drawn from the movable plate without any marked diminution in the original charge. This seems like getting something for nothing, but it involves, each time the plate is lifted, an expenditure of energy to separate the charges against their mutual attraction. This work is in excess of that done against gravity, and the energy of the spark at each discharge represents this excess.

606. The Toepler-Holtz machine. This is the most common form of "influence machine" and can be built to give intense sparks a foot long under all but the worst atmospheric conditions. It consists of two discs, usually of glass, one fixed, while the other, parallel to it and separated by about a centimeter, is made to revolve on a shaft which is turned by the operator. On the back of the fixed plate are two "inductors," aa' , made of paper and tinfoil and indicated by dotted lines in Fig. 38. These act through the glass inductively on metal

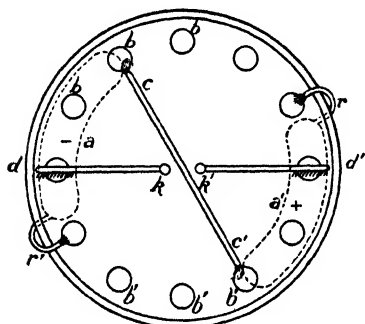


Fig. 38.

buttons, bb' , mounted on the outside of the revolving disc, shown nearest the observer. When given a small initial charge, say negative, a induces two charges on the nearest buttons, b , exactly as the rubber disc acts upon the movable plate of the electrophorus. As b is about to leave the field of a , a brush attached to cc' , the neutralizing rod, carries off the repelled charge, which just balances a similar but unlike one from b' diametrically opposite. Still further separation from a causes the potential of b to rise to a point where it can discharge positive electricity by contact with the replenishing brush r . This carries off a small portion of the button's charge to the positive inductor a' . Finally, when the button reaches d' , it is completely discharged by the points on a collecting comb, which carry the electricity to the discharging knobs, kk' . The points of the comb really act to form a highly concentrated induced charge of opposite sign, which flows across the air gap, neutralizing the charge on the button, and so leaving the knob k' charged with positive electricity like the button. The same process goes on in the lower half of the disc. The inductor a has been replenished, or built up, and the left-hand comb has discharged positive electricity to neutralize the charge on the button b' , leaving its knob negatively charged.

If operated as described above, there will be a rapid succession of very weak though long sparks between the knobs. To make them more intense, such machines have two Leyden jars connected to the discharging rod, as shown in Fig. 39, with their outer coatings connected by a wire. The inner coats acquire opposite charges as the disc revolves, and their potential difference steadily increases as one

risers above and the other falls below zero potential, while the outer coatings remain at zero. Finally the potential difference is sufficient to cause a discharge between the knobs which is much more intense than before, because the quantity discharged from the jars is much greater than that carried by the individual buttons. Naturally, the frequency of the sparks is diminished by this process, for there is only just so much electricity produced, and its accumulation, in order to cause a heavier spark, involves the suppression of many of the weaker discharges.

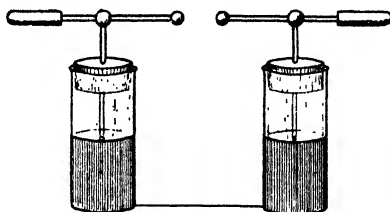


Fig. 39.

607. The Van de Graaff generator.[†] The basic principle of this machine was first used by Lord Kelvin in his famous "water-dropping electrical machine." This involves a continuous and reciprocal induction of moving charges. Positive charges are continually inducing negative charges, and these negative charges induce positive charges which, in turn,

create more negative charges, and so on.

In one form of the Van de Graaff generator, a pulley *H*, shown in Fig. 40, driven by an electric motor, keeps a silk belt in constant motion around a similar pulley *J* inside the hollow sphere *P*. This sphere serves as the positive collector and is mounted on the insulating support *S*. Let us assume a small initial negative charge, due to friction, on the

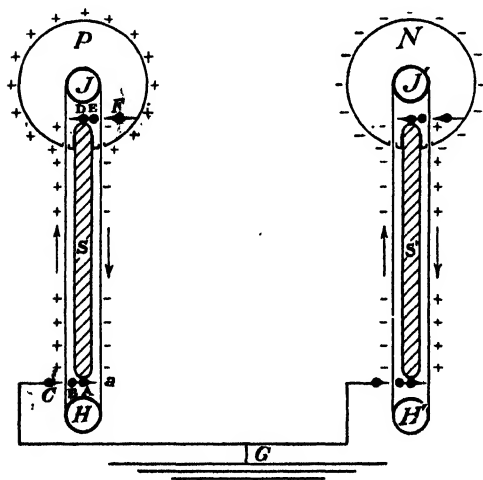


Fig. 40.

belt at *a*. It will discharge to the point *A*, which is connected to the sphere *B* so that the resulting small negative charge on *B*

[†] Described in the *Physical Review* of February 1, 1933, by R. J. Van de Graaff, K. T. Compton, and L. C. VanAtta of the Massachusetts Institute of Technology.

acts inductively on the point C , thus bringing a positive charge from the ground G . This induced charge is "sprayed" from C upon the belt, and is carried upward until it loses its charge to the point D . From D the charge is conducted to the sphere E , which acts inductively on F in the manner just described, drawing negative electricity from the sphere that is left with a positive charge. The induced negative charge sprayed onto the belt is then carried down to a , and the process is thus made continuous.

A similar sequence of events goes on in the other half of the apparatus, but with opposite signs, so that the sphere N acquires a negative charge. In a comparatively short time the potential difference between the spheres builds up to values far higher than have ever before been obtained by electrostatic devices. In the largest of the Van de Graaff generators, the diameter of the spheres is 15 feet, their centers are 35.5 feet above the ground, and they develop around 10 million volts of potential difference.

608. Discharges in general. There are three ordinary ways in which a charged body may lose its electricity: by conduction, by a brush discharge, and in a spark. The latter is said to be disruptive, and involves a sudden breaking down of the dielectric under the intense strain to which it has been subjected. Highly polished spherical knobs with air between them produce the best sparks. The charge is then fairly evenly distributed over their opposite surfaces, whose curvature is not large enough to cause very strong concentration and consequent "brush discharge." Therefore polished knobs do not discharge until the dielectric between them yields. This involves a collapse of the lines of force, which thereby radiate energy into space, as will be seen in another chapter.

A brush discharge occurs at sharp points, or along fine wires of large sectional curvature. In a darkened room this is easily seen as a rosy glow, and it involves a silent discharge of the conductor into the surrounding air. The air becomes highly electrified with the same sign of electricity, and is repelled from the point or wire in a "wind" capable of extinguishing the flame of a candle.

As was explained in Article 597, points may be used to discharge a body from which they project, or may charge it inductively when they project from an uncharged conductor which has been grounded. Lightning rods frequently act in this way during a thunderstorm when no lightning (spark discharge) occurs. At night they may be seen surrounded by the corona of a brush discharge, which is accompanied by a crackling sound when a charged thunder cloud passes

over them. This effect is due to the charge of opposite sign, induced by the cloud in the earth below it, which is discharged into the air by the lightning rod. Such an unloading of electricity may even serve to discharge the cloud sufficiently to prevent lightning from striking a house so protected. Very high tension transmission lines are often surrounded by an invisible corona, which is a similar phenomenon and results in serious line losses, if not prevented.

Finally, discharges may be made by grounding the charged body through a conductor. In this case the discharge is practically instantaneous, and produces what is called a conduction current. No conductor can support a field of force, so the lines representing it collapse, as in the case of a spark. Thus a flow of electricity is associated with the motion of electric charges, and if the amount of electricity to be discharged is limited, the field disappears when the charge becomes neutralized either by grounding or by uniting with an equal quantity of opposite sign.

SUPPLEMENTARY READING

S. P. Thompson, *Elementary Lessons in Electricity and Magnetism* (pp. 44–80), Seventh Edition, Macmillan, 1915.

PROBLEMS

1. A sphere of 3 cm radius is raised to a potential of 81 erg/esu in air. What is the surface density of its charge? *Ans.* 2.15 esu/cm².

2. What is the radius of a sphere immersed in a medium whose dielectric constant is 3, if a charge of 192 esu raises it to a potential of 16 erg/esu? *Ans.* 4 cm.

3. Calculate the capacity of a condenser made of two metal discs of 12 cm diameter and separated by a sheet of average rubber 2 mm thick. (The term "average" refers to the dielectric constant.) *Ans.* 99 units.

4. The condenser in Problem 3 is charged with 330 esu. What is the resulting potential? What is the potential if the thickness of the rubber is increased to 5 mm, the charge being the same? *Ans.* $3\frac{1}{3}$ erg/esu; $8\frac{1}{3}$ erg/esu.

5. An insulated metal sphere of 12 cm radius and charged with +60 esu is connected by a long fine wire to another insulated sphere having a radius of 4 cm and a charge of -36 esu. What is the potential of the system after contact? *Ans.* 1.5 erg/esu.

6. Two spheres of radii 15 cm and 9 cm are charged in air to potentials of +6 and -4 erg/esu respectively. What are their respective energies? *Ans.* 270 ergs; 72 ergs.

7. What are the charges on each of the spheres in Problem 6, if they are brought in contact with each other? *Ans.* +33.75 esu; +20.25 esu.

8. What are the energies represented in Problem 7, and how much is lost in the heat of the partial discharge? *Ans.* 37.97 ergs; 22.78 ergs; 281.25 ergs.

9. A sphere whose radius is 3 cm is brought into the field of another sphere of radius 5 cm which is charged to a potential of 16 erg/esu when their centers are 10 cm apart. The uncharged sphere is then grounded, thus making its potential zero, and later removed to a distance of 120 cm. What are its charge and final potential? *Ans.* -24 esu; $-7\frac{1}{3}$ erg/esu.

10. A condenser is made of two 8×12 cm metal plates separated by a thin layer of air. What is the field intensity between the plates near their centers when they have opposite charges of 160 esu each, assumed to be evenly distributed over their surfaces? What would it be if they were separated by crown glass ($K = 6$)? *Ans.* 20.9 dyne/esu; 3.5 dyne/esu.

CHAPTER 46

Electrodynamics

609. The electric current. The science of electricity in motion, and all related phenomena, are known as **electrodynamics**. In the last chapter we saw that an electrostatic discharge can take place by conduction along a wire. When this occurs, a current is said to flow, and this current will continue as long as there is a difference of potential between the ends of the conductor. In electrostatics, this usually lasts only a very short time, but it may be somewhat prolonged by using a poor conductor, such as a moistened thread. Suppose, for instance, we connect two insulated brass spheres by means of a long cotton thread moistened with salt water, which is then allowed to become nearly dry. If an electroscope is connected to one of the spheres while the other is charged, it will be found that the leaves of the electroscope begin to diverge, quite rapidly at first, and then more and more slowly, reaching a maximum divergence only after quite an appreciable time. This shows that the discharge was most rapid and produced the greatest flow of electricity when the difference of potential between the spheres was greatest, and was complete only when the two ends of the thread reached the same potential. Even very poor conductors cannot long prevent such an equalization, but yield to the strain put upon them, by permitting a current to flow which carries part of the charge on one sphere to the other.

This experiment is the electrostatic equivalent of connecting two tanks, standing at the same level, with a pipe of small diameter, and filling one of them with water. In time the water will come to the same height in both, as the flow, rapid at first, and then more and more gradual, partly empties one and partly fills the other. If the tank originally filled is now emptied, the water flows the other way, bringing both to the same, but lower, level again.

In either case, electric or hydraulic, in order to maintain a steady flow for an indefinite time, it is necessary to keep the original difference of potential constant. This may be effected in both cases by using a pump-like device to take the excess from the lower level and restore it continually to the higher level of the supply. Naturally a supply of energy is needed to make a medium, whether water or electricity,

flow "up hill." But the continuous supply just proposed can do work as it flows down hill again, so the energy expended reappears as heat or in some other form.

610. The electric battery. The "pump" just referred to, by which a constant potential difference may be maintained in order to produce a steady current, is an essential part of any circuit. There are four principal types of "pump." One uses the chemical action of an electric cell, one the electromagnetic action of the generator, one the action of heat on a thermoelectric couple, and a fourth the action of light on the photoelectric cell. For the present we shall consider only the first of these and take its mode of action for granted.

Strictly speaking, the word *battery* means a *number* of "primary cells" connected together, just as we talk about a battery of cannon. In fact, that is the origin of the word. But common usage has inaccurately adopted "battery" as meaning either one or many cells used to produce a constant flow of electricity. The most usual type of battery is the dry cell which has two terminals or poles, and when these are connected by a wire, a current flows along it, producing various characteristic phenomena. The *direction* of flow is really rather meaningless, but it is customary to consider a positive charge as having a higher electrical level than a negative one, and therefore when such charges are connected by a conductor, the flow is thought of as from positive to negative. The motion of the electrons in the metal, on the other hand, is from negative to positive.

In order to determine which is the positive and which the negative pole of the cell, we may use a device designed by Volta consisting of

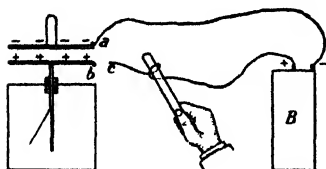


Fig. 41.

an ordinary gold-leaf electroscope having a metal disc in place of a knob. Its upper face is coated with a non-conducting varnish, and a similar plate, also varnished, is fitted with an insulating handle. Now let us connect the cell, as indicated in Fig. 41, in such a way that one pole is connected

to a polished knob at *a*, while the other may be brought in contact with *b* by touching it with the end of the wire *c*. This wire must be carefully insulated during the process, as by wrapping it around a glass rod held in the hand. After the contact has been made, *c* may be withdrawn and then grounded if desired. Now if the upper disc is removed by the insulating handle, the leaf swings out as the charge on the lower disc spreads over it and the fixed plate, used here

in place of another leaf. Two or three cells connected in series give excellent results, though even one cell causes an appreciable divergence of the leaf.

The experiment just described is explained by the fact that the discs form a condenser, one disc becoming positively charged, and the other negatively, though the net result is nearly zero potential as far as the leaf is concerned. However, when the discs are separated, the capacity falls as the distance between them increases (see equation (3), Article 600), and the potential of the lower one rises if it is positive, or falls if it is negative. This is because $V = q/C$, and when q is constant, V varies inversely as C . The rise of potential causes a flow of electricity down into the leaf, which swings out in consequence. The nature of the charge thus acquired may be examined by bringing up an electrified rubber rod which, in the case assumed, will cause the leaf to collapse. This proves that the pole of the battery that charged it was its positive terminal.

611. Rate of flow. When a wire is connected between the positive carbon and negative zinc terminals of a dry cell, a current flows from the former to the latter through the wire, in accordance with the arbitrary assumption that a positive is at a higher potential than a negative charge. This will continue as long as the battery functions like a pump in maintaining the potential difference which the discharge tends to diminish. As in the case of the water analogy, the same current must be flowing through the interior of the cell, but from low to higher levels, since electricity, like water, behaves as an incompressible fluid. Therefore a certain flow, measured in units of the medium per second, in one portion of a closed circuit must involve an equal flow of the medium at every other portion.

The notion of rate of flow, or current, must be clearly understood, for it is not the same thing as velocity. Both in electricity and hydraulics it is possible to vary the character of the flow from point to point. In a river, for instance, we have pools and rapids where the *velocity* is very different, but the *flow*, measured in gallons which pass a given point in a second, is everywhere the same. In currents of electricity there are no varying velocities as in a river, but instead there are varying current densities, depending upon the cross section of the conducting wires. In this way the actual flow is maintained constant all around a closed circuit, which may include a variety of conductors.

612. Effects of electric currents. The most evident effect of an electric current is that it heats the wire which conducts it. A fresh dry cell will heat a short fine wire, connecting its terminals, to redness.

The filaments of incandescent lamps and the coils in electric cooking devices are familiar examples of this important phenomenon. Another effect is chemical, and occurs when a current is passed through certain solutions known as electrolytes, with a resulting decomposition and the formation of different substances at the terminals, or *electrodes*. But from a practical point of view, as well as theoretical, the most important effect of an electric current is its production of a magnetic field surrounding the conducting wire. This phenomenon, which is the connecting link between electricity and magnetism, is the basis of most of the applications of electricity in industry. It includes such great realms as that of the electric motor with its countless uses in the production of mechanical power, of the telegraph, the telephone, both with and without wires, and of numerous other devices in which the magnetic field surrounding a current is an essential feature of the mechanism.

613. Magnetic field around a current. The fundamental fact that electricity in motion produces a magnetic field of force was discovered by the Danish physicist H. C. Oersted in 1820. Until that time electricity, both in static charges and in currents, and magnetism, were regarded as separate phenomena. The experiment by which their interrelationship was established consisted in setting a magnetic

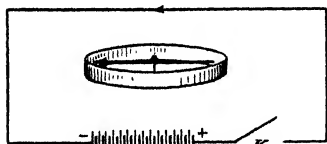


Fig. 42.

compass under a wire in which a current was flowing, as shown in Fig. 42. Currents large enough to affect a needle near a single strand of wire were not common in Oersted's time, so the effect was not as easily obtained as one might imagine. Consequently the scientific

world was much astonished to learn that the needle was deflected whenever the key *K* was closed, and always in the same sense, depending upon the relative direction of the current and the earth's magnetic meridian. If the current flows north over the needle, the latter points westward, more or less, according to the current's strength. Reversing the current makes it point more or less toward the east. These directions are reversed if the wire lies just under the needle instead of over it.

A modification of this experiment consists in placing the wire in a vertical position, in which case the compass needle is always deflected toward tangency with a circle having the wire at its center, as suggested in Fig. 43(a). Here the compasses rest on a board through which the wire passes, with the current flowing up.

If iron filings are sprinkled on the board, as in (b), while it is gently tapped to shake them up, they will arrange themselves in concentric circles. These are quite pronounced close to the wire, but become fainter at increasing distances. Such experiments all go to prove that a conductor carrying a current is surrounded by magnetic lines of force forming closed circles in a plane at right angles to the current direction. Their sense is clockwise if we imagine ourselves looking along the wire in the direction in which the current is flowing. Another useful rule is that of the ordinary right-handed screw, which advances when turned clockwise. If the direction of the current is taken as that of the point of the screw, then the lines of force are directed in the sense of its rotation.

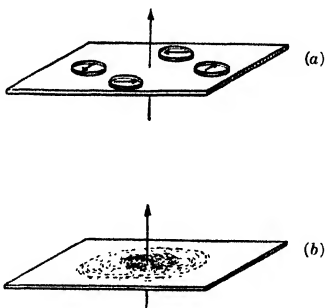


Fig. 43.

614. Biot and Savart's law. Quantitative measurements conducted by these investigators in 1820 demonstrated the fact that the magnetic field in the experiments just described varies inversely as the distance from the wire, provided the wire is long enough to be regarded as of infinite length compared to the radii of the lines of force considered. This relation can be proved in a number of ways, among them being

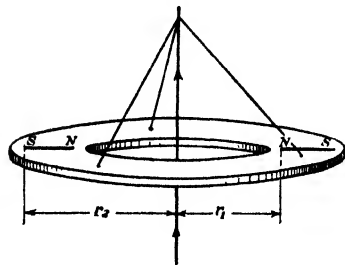


Fig. 44.

the following due to Maxwell: Let a bar magnet, NS , in Fig. 44, be supported on a circular card beside a long vertical wire carrying a current, so that it is free to rotate as a whole around the wire as an axis. Then there is a clockwise torque on N equal to mH_1r_1 , and a counterclockwise torque on S equal to mH_2r_2 , where H_1 and H_2 are the fields caused by the current at the distances r_1 and r_2 from it. But no rotation results from this arrangement; therefore the two moments must be equal, and

$$mH_1r_1 = mH_2r_2,$$

or

$$\frac{H_1}{H_2} = \frac{r_2}{r_1}.$$

Therefore the field at a point due to a constant current in a long straight wire varies inversely as the distance from the wire, as has just been stated.

Experiments with currents of different strengths also prove that H varies directly as the current I , so that we may now write the Biot-Savart law in the form

$$H = \frac{bI}{r}, \quad (1)$$

where b is the constant of proportionality to be determined.

615. Laplace's equation. From the relation established in the last paragraph, Laplace derived an equation which gives the field at a point p distant r centimeters from an infinitesimal portion of a wire dl carrying a current I . When the element is inclined at any angle to the line connecting it with the point, Laplace's equation becomes

$$dH = \frac{kIdl}{r^2} \sin \alpha,$$

where k is a constant and Idl is the current element, as shown in Fig. 45. This shows that H is a maximum when dl is perpendicular to r ,

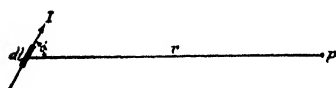


Fig. 45.

and zero when the current element coincides with r . It also shows that the field varies inversely as the square of the distance, instead of the first power as when the wire is straight

and infinitely long. The direction of the field created by the current element Idl at any point p is perpendicular both to dl and to r drawn to that point. In Fig. 45, dl lies in the plane of the paper; therefore H is normal to the paper at p .

616. Magnetic field at the center of a circular current. Laplace's equation, though obtained from the law of Biot and Savart, is still more fundamental, more general, and more useful than its predecessor, and may be used to obtain H in a variety of cases. The most important of them is the field at the center of a single circular "turn" of wire carrying a steady current, as in Fig. 46. In this case all the elements dl are perpendicular to r ; therefore in Laplace's equation $\sin \alpha = 1$, and r , as well as I , is constant. The sum of all the infinitesimal portions, dl , of the turn taken around the circle

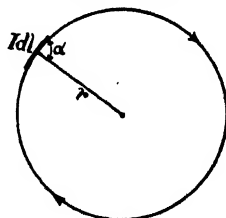


Fig. 46.

is equal to the circumference, $2\pi r$. Therefore the equation becomes

$$H = \frac{kI}{r^2} \times 2\pi r,$$

or

$$H = \frac{2\pi kI}{r}. \quad (1)$$

617. Unit of current. The preceding result is used as a basis on which to define the absolute unit of a current of electricity in the *electromagnetic system*, because it depends upon a measurement of H under comparatively simple conditions. As this unit is the first to be defined that connects the electric current with magnetism, we are permitted a certain freedom of choice, and it has therefore been agreed to set k arbitrarily equal to unity. We may then adopt the following definition: *the absolute electromagnetic unit is one which, flowing in a single circular turn of unit radius, produces a field of 2π oersteds at the center of the circle.* This is evident from the equation, for if k , r , and I are set equal to unity, $H = 2\pi$. As a result of this definition of unit current, the value of H obtained in Article 616 reduces to $H = 2\pi I/r$. When there are N turns of wire wound in a circular coil of negligible cross section, the field at the center is given by

$$H = \frac{2\pi NI}{r}. \quad (1)$$

For practical measurements, the **ampere** has been adopted as the unit of current, so named in honor of the famous French physicist† who discovered many of the magnetic properties of electric currents immediately after Oersted's experiment in 1820. It is defined as one tenth of the absolute unit, or *abampere*. Expressed, then, in amperes, equation (1) becomes

$$H = \frac{2\pi NI}{10r}. \quad (2)$$

The dimensions of current in e.m.u. (electromagnetic units) are obtained from the equation which was used in defining it. We have seen that the dimensions of H are $[M^{\frac{1}{2}}L^{-\frac{1}{2}}T^{-1}]$. Therefore, since $I = Hr/2\pi$, and since 2π has no dimensions, while r has that of a length,

$$[I] = [M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}].$$

Here we have defined the current by its magnetic effect. But we might also have obtained a natural definition as the number of

† A. M. Ampère, 1775–1836.

electrostatic units of charge passing a given point in a second. This is exactly analogous to the definition of the flow of water through a pipe as so many gallons per second. A current of electricity defined in this way comes under the head of the electrostatic system of units, consequently its unit in that system is a current which carries one esu per second past a given section of a conductor. Since charge q has the dimensions $[M^{\frac{1}{2}}L^{\frac{3}{2}}T^{-1}]$, the e.s.u. of current has the dimensions of q/t , or $[M^{\frac{1}{2}}L^{\frac{3}{2}}T^{-2}]$.

We may now ask, what is the relation between these two definitions of current? In other words, how many electrostatic units of charge pass in a second when a unit current defined by its magnetic effect is flowing? This relation may be expressed by

$$i = cI,$$

where i is measured in esu/sec., I is measured in e.m.u. and c is a constant to be determined. Taking the ratio of the dimensions of i and I , we find that c has the dimensions $[LT^{-1}]$. This is a velocity, and when i and I are measured in absolute units, it is found experimentally that $c = 3 \times 10^{10}$ cm/sec. This is numerically equal to the velocity of light! It is no mere coincidence, but a consequence of the fact that light waves are electromagnetic in character.

618. Test of Laplace's equation. Although we have as yet given no proof of this equation, and stated only that it is a consequence of Biot and Savart's observations, it is well to note that it leads to a

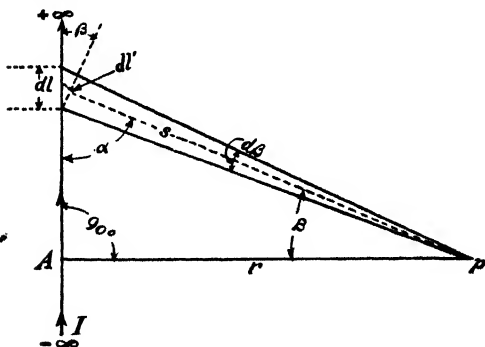


Fig. 47.

calculation of H at the center of a single turn, of many turns, and of coils of various dimensions, so its truth has been verified by countless experiments. We may also test its validity in a rather backhanded way by deriving Biot and Savart's equation for the field surrounding a straight wire carrying a

current, and thus obtain a particular and easily verified case from the general relation. In Fig. 47, let a current I flow in the vertical wire of which one element, dl , is indicated. Let the point p be at a distance r from this wire, and let dl' be the projection of dl on a line normal

to s which connects dl and p . Then by Laplace's equation, the field at p , due to the current element $I dl$, is given by

$$dH = \frac{kI dl \sin \alpha}{s^2}, \quad (1)$$

or, since β and α are complementary angles,

$$dH = \frac{kI dl \cos \beta}{s^2}. \quad (2)$$

But $dl' = dl \cos \beta$ by construction, and the infinitesimal angle $d\beta$ subtended by dl' equals dl'/s . Hence

$$dl \cos \beta = dl' = s d\beta; \text{ also } s = r/\cos \beta. \quad (3)$$

Substituting these values of $dl \cos \beta$ and s from (3) in (2), we obtain

$$dH = \frac{kI \cos \beta d\beta}{r}. \quad (4)$$

Now β varies from 0, when d is at A , to $\pi/2$ radians in either direction, because the wire is supposed infinitely long and r is finite. Between these limits $\cos \beta$ varies between unity and zero, and its average value is known to be $2/\pi$. Therefore, the value of H due to the upper half of the wire may be obtained by taking the entire range of β from zero to 90° , a total of $\pi/2$ radians, and using the average value of its cosine over that range. Then the field intensity at p , caused by both halves of the wire, is twice this value, giving

$$H = \frac{2kI(2/\pi) \times (\pi/2)}{r}$$

or

$$H = 2kI/r.$$

If the current is measured in e.m.u., $k = 1$, and

$$H = \frac{2I}{r}. \quad (5)$$

This is Biot and Savart's law, as already given by equation (1), Article 614, but we have now obtained the value of the constant b , which is shown to be the number 2.

619. Direction of the field. It should be carefully noted that the field at the center of the turn of wire used in defining I is perpendicular to the plane of the circle. The force of 2π dynes is neither an attraction nor a repulsion between the current and the imaginary pole. But if a single pole could be placed at the center, it would move at

right angles to the plane of the wire shown in Fig. 48. The direction is vertically upward if the current is flowing counterclockwise when seen from above. The direction is vertically downward if the current flows in the opposite sense.

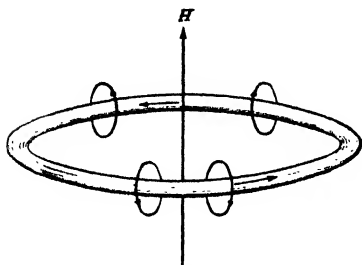


Fig. 48.

This follows from the fact that the force is due to a field represented by lines of force forming closed loops around every element of the current, as suggested by a few indicated above. They all unite to produce the field H perpendicular to the circuit at its center.

If the current flows in a straight conductor instead of in a circle, an isolated pole placed at a distance r from it would travel around it in perfect circles of radius r , the wire being the axis. Thus the relations between current and field have been exactly reversed, and as before, there is no attraction between pole and wire as a result of the circular field around the latter. It will also be seen that in both the cases we have discussed, the field and current *interlink* each other, and this interlinking of electricity in motion with the magnetic flux is of the utmost significance in electromagnetic theory.

620. Rowland's experiment. Since an electric current is really a stream of electrically charged *particles*, we should expect that charged *bodies* in rapid motion ought to behave like a current and produce a magnetic field. This was found to be the case by H. A. Rowland, an illustrious American physicist (1848–1901). He used two non-conducting discs (Fig. 49) which faced each other and could be rotated in opposite directions at a very high speed. The two opposed faces were coated with gold leaf and given unlike charges, negative and positive. When the discs were spinning, a magnetized needle placed between their centers was deflected as it would be by a current flowing in the sense of rotation of the positively charged disc. If only one disc were rotated, the effect was half as great. This classic experiment proved that the charges were carried around with the conducting surface, and were equivalent to electric currents flowing in the same sense as that of the positive charge.

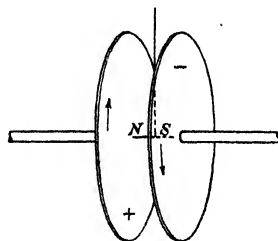


Fig. 49.

It should also be noted that the experimental difficulty of Rowland's experiment was very great, because the effect is vastly smaller than that caused by an ordinary current flowing in a wire. Even small currents carry what would be enormous electrostatic charges.

621. Action of a magnetic field on an electric current. In Oersted's classic experiment a current caused a magnetized needle to turn. But in accordance with Newton's third law, there is an equal and opposite reaction tending to move the wire. Stated in broader terms, *a magnetic field exerts a force upon a wire carrying a current.* As we have seen, such a wire is surrounded by con-

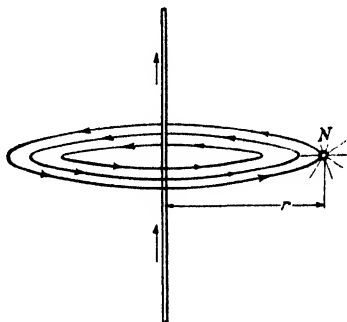


Fig. 50.

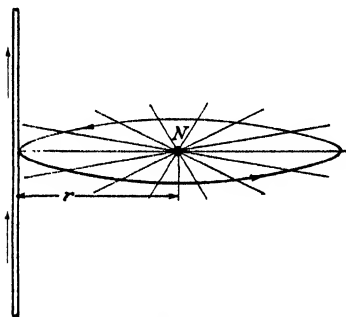


Fig. 51.

centric lines of force, and a north pole, if left free to move, would follow a line of force and rotate continually around the wire in a circle of a constant radius without being either attracted or repelled by the wire, as indicated in Fig. 50. On the other hand, if the wire were free to move so that its axis was always parallel to itself, as in Fig. 51, it would revolve in a circle of radius r around the pole and therefore cut the

horizontal component of the pole's field at right angles. In either case the relative motion is the same, the only difference being that in one case the wire and in the other the pole is fixed.

If the lines of force are straight and parallel to each other, as in a uniform field like that of the earth, the wire still moves at right angles to the lines of force. Its path is therefore a circle of infinite radius, which is a straight line. If the wire whose section is shown at A in

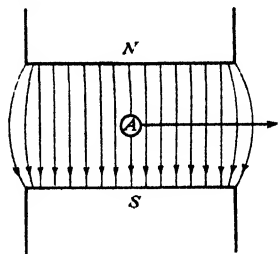


Fig. 52.

Fig. 52 carries a current directed upward from the plane of the paper, it will be urged to the right, and move across the uniform field pro-

duced near the center of the space between the wide poles NS . Thus it is neither attracted nor repelled by these poles, but moves *across* their field at right angles to it. This is a very important fact, and differentiates electromagnetic forces from any other kind. The resulting force is simultaneously at right angles to its two causes, field and current, while these are perpendicular to each other. Thus the three vectors involved, like the three dimensions of space, are perpendicular each to each.

The magnitude of the force with which a uniform field acts upon a wire carrying a current may be found as follows: Consider a circular turn of radius r carrying a current of I e.m.u. Then H_c at the center is given by equation (1) of article 617, where $N = 1$, or $H_c = 2\pi I/r$. If a pole of strength m is placed at the center of the turn, the force acting on it is given by

$$F = 2\pi Im/r. \quad (1)$$

Multiplying and dividing (1) by r , we have $F = 2\pi rIm/r^2$. But $2\pi r$ is the length of the wire; therefore

$$F = Ilm/r^2. \quad (2)$$

Now the force exerted by the current on the pole must be equal and opposite to the force exerted by the pole on the wire. This force is caused by the field H of the pole at the distance r . So $H = m/r^2$ at the wire, and is everywhere normal to it. Then setting $m/r^2 = H$ in (2), we have

$$F = HIl \text{ dynes,}$$

or

$$F = HIl/10 \text{ dynes,} \quad (3)$$

if I is in amperes.

This fundamental relation has been derived for the special case of a circular turn, but it is obviously true for any small portion l' of the turn, and by increasing r , we may make l' as straight as we please. We are then justified in applying (3) to a straight wire of length l , at right angles to a uniform field H , as shown in Fig. 52.

622. Unit of quantity. When a current I flows through a circuit for a time t , the product It means the **quantity** transported by the current in that time. Its absolute e.m.u. is one abampere-second. This quantity will be designated by Q , instead of q , as in electrostatics, and it has different dimensions. Its practical unit has an especial name, the **coulomb**, after the French savant already mentioned. When I is measured in amperes, and t in seconds, their product gives the quantity in coulombs; thus one coulomb equals one ampere-second. As the

ampere is one tenth of the absolute unit, the coulomb is also one tenth of the absolute unit of quantity. It is analogous to the gallon in hydraulics, but the ampere has no analogue, since the flow of water is always expressed as so many gallons or cubic feet per second. Thus the single word ampere saves us from saying "coulombs per second." The dimensions of quantity in the electromagnetic system are those of a current multiplied by time, or

$$\begin{aligned}[Q] &= [M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1} \times T] \\ &= [M^{\frac{1}{2}}L^{\frac{1}{2}}].\end{aligned}$$

The ratio of the dimensions of Q in e.s.u. to those of q in e.m.u. is 3×10^{10} cm/sec., as in the case of the currents (Article 617). This is obvious because in both cases $Q = It$. Thus one coulomb (10^{-1} emu) equals 3×10^9 csu.

623. Joule's law. The preceding discussion is all based on the experimental evidence relating to a single phenomenon, namely, the interlinking of the electric current and the magnetic flux. We now come to a second experimental fact in the realm of moving electricity which connects the current with thermal energy. When a current flows along a wire, it heats the conducting metal. In 1841, J. P. Joule, an English scientist, undertook quantitative measurements of this already well-known fact. He showed that the heat evolved is directly proportional to the square of the current in the wire, and directly as the time during which it flows. This fact, known as **Joule's law**, may be expressed as

$$W = I^2Rt,$$

where W is the heat evolved and R is the constant of proportionality, which differs according to the length, cross section, and material of the conductor. This constant is known as the **resistance** of the wire.

624. Units of resistance. In order to obtain the absolute unit of resistance, we must choose the erg to measure W ; then, taking the second and the absolute unit of current for the other quantities, R is reduced to unity in the absolute system of electromagnetic units. This results in an excessively small resistance, so the practical unit is taken 10^9 times as large, and is called the **ohm**, after the German physicist G. S. Ohm (1787–1854). The choice of 10^9 as the factor is not wholly arbitrary, for as the ampere is one tenth of the absolute unit, this expresses W in units of 10^7 ergs, that is, in joules. Thus we may define the ohm as *that resistance which develops a joule of heat per second when one ampere flows through it.*

If, however, we wish to obtain the heat developed in calories, it is necessary to introduce Joule's equivalent into the equation, which then reads

$$W \text{ (calories)} = 0.239 I^2 R t.$$

The time rate at which electrical energy is converted into heat expressed in joules per second is $I^2 R t/t$. This is $I^2 R$ joules per second, or watts. This power is always wasted except when the resistance is purposely used to heat something, as the filament of an incandescent lamp or an electric heater. Consequently, in all other uses of electricity the apparatus is designed to make the $I^2 R$ loss as small as possible.

From Joule's equation we may obtain the dimensions of R , because those of W and I are known. Thus, substituting $[W] = [ML^2T^{-2}]$ and $[I] = [M^{1/2}L^{1/2}T^{-1}]$, we have

$$[R] = [W/I^2t] = \frac{[ML^2T^{-2}]}{[M^1L^1T^{-1}]^2 \times [T]},$$

or

$$[R] = [LT^{-1}].$$

Thus the absolute unit of resistance, or **abohm**, is one cm/sec. in e.m.u.

Another definition of the ohm has been adopted by an international scientific congress and is called the **international ohm**. This unit, while not a physical quantity in the true meaning of the word, is very convenient for manufacturers of standard resistances, and has been legalized by the government. It is the resistance of a column of mercury contained in a tube of uniform bore having a diameter of 1.1284 mm and a length of 106.300 cm when the temperature is 0° C. Or, as given by law, the bore is not specified, but the mass of mercury must be 14.4521 grams at the temperature of melting ice.

625. Potential difference and electromotive force. Joule's equation may be rearranged to bring in the quantity Q as follows:

$$W = I^2 R t = IR (It).$$

Then substituting Q for It , we obtain

$$W = IRQ. \quad (1)$$

This is the work done when Q units of electricity pass through the circuit. But whenever it requires work to move a quantity of anything against a force of any kind, a difference of potential exists between any two points along its path. As usual this potential difference is measured in work per unit quantity. In the circuit shown in

Fig. 53, let V_1 be the potential at the point A , and V_2 the potential at B . Then their difference, $V_1 - V_2$, equals the work done in driving unit quantity along the wire from one point to the other. That is, $V_1 - V_2 = W/Q$. But from (1) above, $W/Q = IR$. This product IR is taken as a new unit whose symbol is E (or ΔV), and is defined by the identities $E = \Delta V = V_1 - V_2 = IR = W/Q$. Thus E (or ΔV) is a potential difference, and is measured by work per unit quantity.

The quantity E may also be used to denote the difference of potential developed by a battery, generator, and so forth, which cause the flow of current. In this case it is called *electromotive force*. The name is unfortunate, for it is not measured in terms of force, but of *work* per unit quantity.

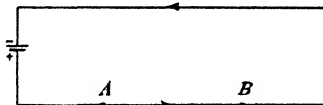


Fig. 53.

626. Units and dimensions of potential difference. *The absolute unit of potential difference (called **abvolt**) is obtained when an absolute unit of current flows between two points in a circuit having an absolute unit of resistance between them.* If I is expressed in amperes (10^{-1} abampere) and R in ohms (10^9 abohm), then since $E = IR$, unit E must be 10^8 times its absolute unit. This is the practical unit, and is named the **volt**, after the Italian, Alessandro Volta, who discovered the “voltaic cell” in 1800. The dimensions of potential difference in e.m.u. are obtained from those of I and R already found, so that

$$\begin{aligned}[E] &= [M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}] \times [LT^{-1}] \\ &= [M^{\frac{1}{2}}L^{\frac{3}{2}}T^{-2}].\end{aligned}$$

We might also arrive at this conclusion in the same way as in electrostatics, by defining the volt as a joule/coulomb, because it measures potential difference and therefore is work/quantity. But as shown in Article 622, 1 coulomb = 3×10^9 esu, so a volt equals 10^7 ergs/ 3×10^9 esu = $1/300$ erg/esu. Or 300 volts = 1 erg/esu, where erg/esu is the absolute unit of potential difference in e.s.u.

In e.m.u., 1 abvolt = 1 erg per abampere-second or 1 erg/emu, so 1 volt = 10^8 erg/emu.

627. Ohm’s law. The identity $E = IR$ may be written $I = E/R$. In this form it is known as Ohm’s law, and was discovered as an independent experimental fact first formulated by Ohm in 1828. This is not, however, a necessary experiment if Joule’s law is accepted; but because of its practical value and the ease with which it is verified,

it is of the greatest importance. This law may be stated in words thus: *The current in a conductor is equal to the difference of potential between any two points divided by the resistance between them.* Or *amperes = volts/ohms.*

628. Specific resistance and conductivity. It has been experimentally determined that the resistance of a conductor of uniform constitution and cross section varies directly as its length and inversely as the area of its section; therefore

$$R = \frac{\rho l}{A}, \quad (1)$$

where ρ is the constant of proportionality, l is the length in centimeters, and A the cross section in square centimeters. The constant depends for its value on the nature of the conductor, and not upon its dimensions. It is called the **specific resistance**, or **resistivity**. If the conductor is a unit cube, l and A are each equal to unity, and $R = \rho$. Therefore the resistivity of a conducting material may be defined as *the resistance of a bar of unit section per unit length*, and its dimensions are obviously those of an ohm-centimeter.

The reciprocal of the resistance of a conductor is called its **conductance**, and the reciprocal of its resistivity is known as its **conductivity**. Conductance, like resistance, depends upon both the nature of the conductor and its dimensions. Its unit is the **mho** (ohm spelled backwards). Conductivity, like resistivity, is independent of the conductor's dimensions. Its unit has no name, but is usually expressed by the letter k .

The conductivity of a metal has an important bearing on current density in a conductor. Current density is defined as I/A , where A is the cross section of a conducting bar or wire. Thus, taking ΔV as the potential difference between two points on a conductor at a distance l from each other, we obtain $I = \Delta V/R$. But by definition $R = \rho l/A$; therefore

$$I = \frac{\Delta V}{R} = \frac{A \Delta V}{\rho l}.$$

Dividing by A , we have
$$\frac{I}{A} = \frac{1}{\rho} \left(\frac{\Delta V}{l} \right) = k \left(\frac{dV}{dl} \right), \quad (2)$$

where dV/dl is the potential gradient at any point in the conductor. This was shown for the corresponding case in electrostatics by equation (2), Article 588. We may now state that conductivity is the ratio of current density to potential gradient along a conductor.

The following table gives the resistivities and temperature coefficients (defined in Article 634) at 18° C of some of the more common metals and alloys.

Substance	Resistivity (at 18° C)	Temperature Coefficient
Aluminum.....	2.94×10^{-6}	38×10^{-4}
Bismuth.....	$119. \times 10^{-6}$	42×10^{-4}
Copper (drawn).....	1.78×10^{-6}	38.8×10^{-4}
Gold.....	2.42×10^{-6}	36.5×10^{-4}
Iron (wrought).....	13.9×10^{-6}	62×10^{-4}
Iron (cast).....	74.4×10^{-6}
Steel (soft).....	15.9×10^{-6}	42.3×10^{-4}
Steel (glass hard).....	45.7×10^{-6}	16.1×10^{-4}
Lead.....	20.8×10^{-6}	43×10^{-4}
Mercury.....	95.6×10^{-6}	7.2×10^{-4}
Platinum.....	11.0×10^{-6}	37×10^{-4}
Silver.....	1.66×10^{-6}	38×10^{-4}
Tin.....	11.3×10^{-6}	36.5×10^{-4}
Zinc.....	6.1×10^{-6}	36.5×10^{-4}
Brass.....	6.6×10^{-6}	10×10^{-4}
German Silver.....	33.0×10^{-6}	$2.3 \text{ to } 6 \times 10^{-4}$

629. Electromagnetic energy and power. Joule's equation may be transposed by substituting $E = IR$ so as to read

$$W = EI t,$$

or
$$\frac{W}{t} = EI = P.$$

Thus we see that the product of the current and the potential difference in a circuit gives the rate of production of energy which is measured in joules per second or *watts*, provided E is expressed in volts and I in amperes. Since R has been eliminated, we may use this relation in calculating the power developed in a circuit, or portion thereof, without knowing what causes the expenditure of energy. As will be seen later, there are other ways in which an electric current does work besides developing heat. Whereas the equation $I^2 R = W/t$ measures only the rate at which heat is evolved, EI measures the *total power*. When resistance is the only consideration and Ohm's law holds true, EI , the power expended, equals $I^2 R$. But when other work is being done, such as the mechanical work of an electric motor, this is no longer true. Then $I^2 R$ represents only part of the input, and is therefore less than EI .

SUPPLEMENTARY READING

A. Zeleny, *Elements of Electricity* (Chapters 8, 9, 10), McGraw-Hill, 1930.

PROBLEMS

1. What is the field strength at a point 8 cm from a long straight wire carrying a current of 48 amperes? *Ans.* 1.2 oersted.

2. How many turns are needed in a coil of negligible section and 15 cm radius to create a field of 6 oersteds at its center, when a current of 3 amperes flows through it? *Ans.* 48 turns, nearly.

3. A straight wire 60 cm long is at right angles to a field of 12 oersteds. If a current of 8 amperes flows through it, what is the force pushing it across the field? *Ans.* 576 dynes.

4. What is the potential difference between the ends of a wire of 16 ohms resistance, when 42 coulombs flow through it in 6 seconds? *Ans.* 112 volts.

5. How many calories are produced in 15 minutes by a current of 4 amperes in a coil of wire whose resistance is 12 ohms? *Ans.* 41,299 calories.

6. A current of 5 amperes flows through a wire for 8 seconds, and develops 10 joules of heat energy. What is the resistance of the wire? *Ans.* 0.05 ohm.

7. Calculate the joules per minute developed in coils of 10 and 20 ohms resistance under an impressed e.m.f. of 110 volts. *Ans.* 72,600; 36,300 joules per minute.

8. How many coulombs are needed to develop 500 calories with an e.m.f. of 110 volts? *Ans.* 19 coulombs nearly.

9. Calculate the resistance of a copper wire at 18°C , if its length is 24 m and its cross section is 0.3 mm^2 . *Ans.* 1.424 ohms.

10. What is the length of german silver wire which should be used in a heating element designed to develop heat at the rate of 55 watts on a 110 volt circuit, if its cross section is 0.2 mm^2 , and $\rho = 33 \times 10^{-6}$? *Ans.* 133.3 m.

11. A copper wire has a diameter of 3 mm, and carries a current of 12 amperes. What is the potential gradient? (Use equation (2) of Article 628.) *Ans.* 3×10^{-4} volt/cm.

CHAPTER 47

The Electric Current

630. Circuits. In order that a difference of potential may produce a steady flow of electricity or continuous current, there must be a complete circuit closed upon itself. In the simplest case this consists of a wire connecting the two terminals of a battery. The battery itself forms a portion of the circuit, and exactly the same current flows between its plates that is to be found at every point of the wire. In short, as has already been stated, electricity in motion behaves like an incompressible fluid which a pump causes to circulate through a closed pipe. But there are many possible circuits more complicated than this one, where it is not so evident just how such a current will behave in the different parts of the system. In order to deal with all such cases we make use of two laws formulated by G. R. Kirchhoff (1824–1887), a German physicist. These laws enable us to solve problems when the current is subdivided in very complicated systems of conductors and cells.

631. Kirchhoff's first law. This law states that *the algebraic sum of all the currents which meet at a common point is equal to zero*. Thus in Fig. 54 the currents i_1 and i_2 are arriving at the point P , while i_3 , i_4 , and i_5 are leaving it. If this were a flow of water in pipes arriving at a common junction point, we should unhesitatingly write $i_1 + i_2 = i_3 + i_4 + i_5$. This is also the case with the electric current, but it is customary to set all the terms on one side of the equation, with minus signs before those which are flowing *away* from the junction. Thus we write $i_1 + i_2 + i_3 + i_4 + i_5 = 0$, or more simply,

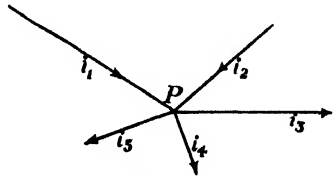


Fig. 54.

$$\sum_n i = 0, \quad (1)$$

which means that the algebraic sum of the n currents is equal to zero.

632. Kirchhoff's second law. This law applies to any one mesh in a network of currents, where there may or may not be sources of electromotive force included in the particular mesh. This law states that *the algebraic sum of all the IR potential differences around any closed*

circuit is equal to the algebraic sum of the electromotive forces in that circuit. Here again we should expect the same thing in hydraulics, for a network of canals and rivers with canal locks at certain points is a close analogue of the network of currents as represented in Fig. 55.

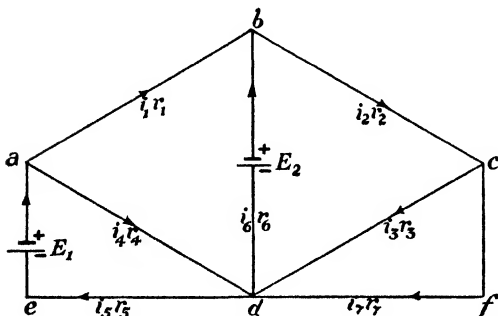


Fig. 55.

The locks are similar to batteries, and the fall or rise of level in going down or up stream corresponds to changes in electrical potential due to resistance. It is self-evident that a canal boat which traverses a series of canals and rivers, finally returning to its starting

point, must have descended as many feet in level as it has risen. So in circuit $abcda$, where there are no batteries, there is a fall of potential in going with the current from a through b and c to d , and then a rise in going from d to a against the stream. Hence, since in general, potential difference equals IR , we may write

$$i_1r_1 + i_2r_2 + i_3r_3 - i_4r_4 = 0,$$

or in general,

$$\sum_n ir = 0 \quad (1)$$

around a closed circuit which contains no sources of electromotive force such as a battery. The negative sign indicates that the potential is rising because we are going *against* the current, and the positive sign indicates falling potential as we move *with* the current.

If a cell of voltage E is included in the circuit, as in $bcd b$, the level changes abruptly at that point, and the equation becomes

$$i_2r_2 + i_3r_3 + i_6r_6 = E_2,$$

or in general,

$$\sum_n ir = E. \quad (2)$$

If there is more than one cell in the circuit, as around $abdea$, then we should write

$$\sum_n ir = \sum_m E, \quad (3)$$

where $\sum_m E$ is the algebraic sum of m electromotive forces acting within the mesh. In the case of the circuit $abdea$, $\sum E$ equals $-E_2 + E_1$, because the potential falls in going through a cell *against* its e.m.f. and rises going *with* it.

It should be noted that r_b and r_s refer to the *entire* resistance of the paths bd and dea , part of which, however, is inside the cell in the form of "internal resistance," whose effects will be considered later.

In the network of Fig. 55 there are four junction points and ten possible closed paths; hence we could write fourteen equations by using both laws. In such problems the resistances and voltages are usually known, so that a complete solution of all seven currents may be made by picking seven suitable equations from among the fourteen.

633. Series and parallel circuits. When two or more resistances are connected end to end, the total resistance is equal to the arithmetical sum of them all. This rather obvious fact may be rigorously proved by using Kirchhoff's second law, and applying it to the circuit shown in Fig. 56, where there is but one current. Then

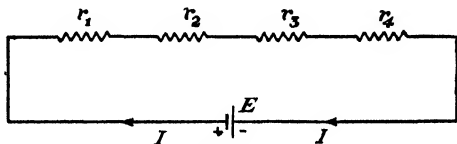


Fig. 56.

$$Ir_1 + Ir_2 + Ir_3 + Ir_4 = E. \quad (1)$$

But according to Ohm's law, $I = E/R$, or

$$R = E/I. \quad (2)$$

Hence, dividing (1) by I we obtain

$$r_1 + r_2 + r_3 + r_4 = E/I = R. \quad (3)$$

When two or more resistances are so connected that the main current divides between them, as shown in Fig. 57, then from Kirchhoff's first law

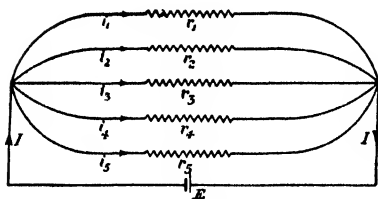


Fig. 57.

$$I = i_1 + i_2 + i_3 + i_4 + i_5, \quad (1)$$

and from Kirchhoff's second law we have for each of the five circuits

$$E = i_1 r_1 = i_2 r_2 = i_3 r_3 = i_4 r_4 = i_5 r_5.$$

Therefore, substituting for each i in (1) above, we obtain

$$I = \frac{E}{r_1} + \frac{E}{r_2} + \frac{E}{r_3} + \frac{E}{r_4} + \frac{E}{r_5}, \quad (2)$$

or

$$I = E \left(\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} + \frac{1}{r_4} + \frac{1}{r_5} \right). \quad (3)$$

But by Ohm's law, $I = E/R$; therefore the total resistance R is equal to the reciprocal of the parenthesis. If there are only two resistances, this becomes $(r_1 r_2)/(r_1 + r_2)$; if there are three, $R = (r_1 r_2 r_3)/(r_1 r_2 + r_1 r_3 + r_2 r_3)$, and so on to any number of branches. This conclusion may be summed up briefly in the rule: *To obtain the*

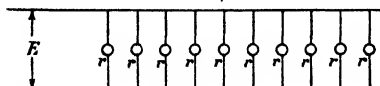


Fig. 58.

total resistance of parallel circuits, add the conductances and invert.

It is convenient to consider the very usual case where a number of equal resistances are in parallel across a constant potential, as in the case of lamps in lighting circuits, shown in Fig. 58. Then since $r_1 = r_2 = r_3 = \dots r_n$,

$$\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} + \dots \frac{1}{r_n} = \frac{n}{r},$$

so that $R = r/n$, and the total resistance of twenty lamps, for instance, is but one twentieth that of a single lamp. From this it follows that the more lamps we turn on in our houses the lower is the resistance to the flow and the greater the current becomes, with a consequent increase in the amount of electrical energy consumed. If no lamps are turned on, there is *no circuit*, and the resistance between the "leads," as they are called, is infinite, which is quite different from *no resistance*. The latter is possible only with a "dead short circuit" between the wires.

Still another useful consideration is the fact that when a current divides between two or more branches in parallel, the currents in the branches are proportional to each other inversely as the resistances. This may be shown as follows: In the branches, $i_1 = E/r_1$, $i_2 = E/r_2$, $i_3 = E/r_3$, and so on. But as E is the same for all, the quotients of these equations give us $i_1:i_2:i_3:\dots i_n :: 1/r_1:1/r_2:1/r_3:\dots 1/r_n$. If there are only two resistances, $i_1/i_2 = r_2/r_1$, which is a frequently used relation.

634. Temperature and resistance. The resistance of all pure metals increases with the temperature in much the same way that the volume of gases increases when they are heated at constant pressure. There is a nearly constant increase in resistance for equal temperature intervals. Therefore we may express the resistance at any temperature in terms of that at 0°C by the familiar equation $R_t = R_0 (1 + \alpha t)$, where α (as with thermal expansion) is the temperature rate of proportional increase in resistance. This equation may also be written $\alpha = \frac{R_t - R_0}{R_0 t}$, which becomes $\alpha = \frac{dR}{R dt}$ for infini-

tesimal changes in the temperature, and must be used in this form when α is not constant.

The influence of temperature on resistance is a consequence of the thermal agitation of the molecules. Their motion tends to impede the circulation of the free electrons by increasing the probability of collision, and consequently the current caused by a given e.m.f. is diminished.

In the case of pure copper and platinum, α is approximately 0.004, which is not far from the values of the expansion coefficient of gases. Its value for iron is 0.0055, and for nickel, 0.006. Iron behaves irregularly at 800° C, the point of recalescence, just as it does in a magnetic field. The value of α rises rapidly to 0.018, and then near 850° it comes back to its former value of 0.0055.

The temperature coefficient of alloys differs very much from 0.004, and in some of them, like manganin, an alloy of copper, manganese, and nickel, α is only 0.00002, so that the effect of heat on resistance is negligible at the temperatures to which it is ordinarily subjected. Manganin is therefore used, with other alloys of the same type, in the manufacture of resistance standards. Carbon and the oxides of the alkaline earths such as cerium, on the other hand, have negative coefficients, so that R falls when they are heated. A rod made of such oxides is practically a nonconductor at ordinary temperatures, but when heated by some outside source, such as a flame or electric coil, it begins to conduct at red heat. If a current is then sent through the rod, the Joule heat makes it incandescent and keeps it glowing as long as the current flows. This principle was used by Nernst in the "Nernst lamp," formerly much in vogue.

The reason why some nonconductors begin to conduct at high temperatures may be the production of free electrons caused by thermal agitation of the molecules. Other nonconductors undoubtedly become ionized, and conduct by the process known as *electrolysis* (Article 644), which occurs in many solutions called *electrolytes*. Electrolytes also have negative temperature coefficients and conduct better when they are warmed.

635. Superconductivity. Some metals experience a very abrupt fall in resistance at temperatures within a few degrees of the absolute zero. Among these are tin, lead, mercury, and thallium. The change occurs at 7°26 K in lead, at 4°12 K in mercury, at 3°69 K in tin, and at 2°38 K in thallium. Until somewhere near this point, the decrease in resistance goes on about as in other metals, and the resistance-temperature curve shows a tendency to flatten out as if toward a

finite resistance at the absolute zero. But then the resistance of superconductors begins to fall much more rapidly until, at a sharply defined temperature, it suddenly drops to less than a ten-billionth part of its usual value. In the case of thallium, for instance, the resistivity at 0°C is 17.6×10^{-6} ohm, at 90°K it is 4.08×10^{-6} , and at 5°K it is 0.4×10^{-6} , but at 2.38°K it drops practically to zero.

When there is no resistance in a conductor there is no lost energy when a current flows, and no Joule heat is developed. Therefore a current once started keeps on flowing. This was tested by Kamerlingh Onnes, who immersed a closed coil of lead wire in liquid helium at about 4°K and placed it between the poles of an electromagnet. A current was then induced in the lead coil by gradually cutting out the current in the coils of the magnet. The strength of the current thus produced decreased so gradually that at the end of four days it still had two thirds of its initial value. To explain this phenomenon, it has been suggested that as the metal cools and shrinks, the atoms crowd closer and closer together, and at length electrons usually held close to the nuclei become free to pass from atom to atom in "some sort of chain gang motion."[†] But the details of what happens to a metal when it passes into the superconductivity state are not at all clearly understood in spite of a great amount of study by many physicists.

636. The bolometer and resistance pyrometer. The bolometer is a very delicate device for comparing and measuring minute quantities of thermal radiation. It was invented and used with great success by Professor Langley, an American astronomer, in his classic study of the distribution of energy in the infrared region of the solar spectrum. The apparatus consists of blackened platinum strips 0.0005 mm in thickness, or of very fine wires. These are placed parallel to the slit of a spectrometer fitted with a rock-salt prism, and in the path of the refracted beam. The strip absorbs the radiation falling on it, its temperature rises, and its resistance is increased. This change of resistance is readily observed, and serves as an index of the amount of energy received in the particular part of the spectrum where the strip is located.

The resistance pyrometer operates in a similar manner, but instead of being used to record very small quantities of heat, its function is to measure temperatures beyond the range of a thermometer. It is essentially a resistance coil made from some refractory metal like

[†] E. L. Hill, "Superconductivity in Metals," *Review of Scientific Instruments*, January, 1933.

platinum whose melting point is sufficiently above the temperatures it is designed to measure. The change in resistance resulting from a rise of temperature is measured, and the corresponding temperature is obtained from a knowledge of its temperature coefficient, or from a curve plotted from its actual performance at known temperatures. Pyrometers of this type may be used to measure temperatures up to about 1000° C.

637. Electric heating. As has already been shown, when a current flows through a conductor, the latter is heated so that I^2Rt joules of heat energy are produced. If we multiply this expression by the reciprocal of Joule's equivalent, we obtain the heat in calories as

$$W = 0.239 I^2Rt,$$

when a current I , expressed in amperes, flows for t seconds through the resistance R in ohms. This unavoidable conversion of electrical energy into heat becomes of practical value in electric cooking, heating, and lighting.

In the conversion of electrical into thermal energy, there is no waste, and the process is always 100 per cent efficient, so that one heating device is just like another as regards the number of calories produced per second by a given number of watts. But the resulting temperature depends upon the design of the apparatus, and much can be accomplished to prevent loss of heat by unnecessary radiation, and in concentrating the heat where it is most useful. The requirements of a heater for cooking, or for warming a room, are given by the number of joules it is desired to develop per second, and the line voltage. Hence, transforming Joule's equation by substituting $I = E/R$, we obtain W (in joules) = E^2t/R , whence the required resistance is readily obtained. Evidently for a fixed E , a low resistance results in a greater evolution of heat than a high one.

638. Further applications of electric heating. Electric furnaces, capable of melting copper, iron, and other metals, are made of coils of a refractory metal such as tungsten or platinum surrounded by some insulating material like porcelain or asbestos. In this way temperatures as high as 1500° C may be obtained.

If still higher temperatures are needed, the electric arc between carbon terminals is employed. Substances placed between the carbons may be heated to 3500° C. All elements have melting points below this temperature, including even tungsten, which melts at 3395° C. There is therefore no known element which cannot be melted in the electric arc.

Iron and other metals are welded in several ways. In butt joints, the two ends are pressed together and a very heavy current is sent through the junction, heating it to welding temperature. Lap joints between metal plates are made by "spot welding." In this process the overlapping plates are placed between two electrodes which press them together. Then a current of several thousand amperes flows across the overlapping plates so that the *spot* between the electrodes is heated to redness, and the plates cohere as if they had been riveted.

639. Arc lights. As in electric furnaces, the arc light depends upon heat created by an electric current. In the carbon arc, the highest temperature, and consequently the most intense light, is in the "crater" of the positive carbon, which reaches 3500°C under ordinary conditions. When surrounded by carbon dioxide under high pressure, its temperature may exceed 6000°C . The negative carbon also glows brightly, but normally reaches only 2700°C . Some light also comes from the arc itself. But most of the light is emitted from the crater, in the direction shown in Fig. 59.

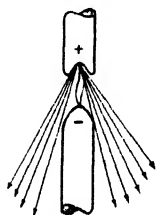


Fig. 59.

The heat developed by the arc is due only in part to the resistance of the carbons. The high temperature of the crater is caused mainly by electrons that are shot out from the negative terminal and strike the positive carbon with great violence. This results in a production of about a candle power of light per watt between carbon terminals, while an arc between a rod of copper and one of magnetite furnishes light at the rate of a candle power for every half watt of power expended.

640. Incandescent lamps. Nearly all commercial electric lights today are of this type. They depend upon the principle that a current of electricity passing through a wire heats it to incandescence. The light produced does not differ from that which would result from heating the same filament in a Bunsen flame, and so is wholly thermal in its origin.

The earliest incandescent lamps had a carbon filament enclosed in a bulb from which most of the air had been exhausted. This precaution is necessary to prevent the rapid oxidation of the filament which would otherwise take place. Later the carbon filament was replaced by tantalum and tungsten. These refractory metals can stand a very high temperature (2000°C) without sublimation, which blackens the bulb and wastes away the filament. The possibility of using higher temperatures with metal filaments results in much more light for the

same expenditure of energy. Then the proportion of the energy converted into visible light as compared to the entire amount, including the long waves of heat, increases very rapidly as the temperature rises. The wave length of maximum energy emitted by tungsten lamps, as calculated by Wien's law, is 1.27 microns. This is not far below the visible red. Still higher temperatures are made possible by introducing into the bulb an inert gas like nitrogen or argon, at about one-third atmospheric pressure when cold. When the lamp gets hot, the pressure rises to about an atmosphere, and so prevents the sublimation of the metal which would occur in a vacuum at the temperatures employed.

As the actual value of the useful part of the light emitted by a lamp cannot readily be found in terms of watts, it is customary to rate its efficiency not as a per cent, but as so many candle power per watt. Ordinary tungsten lamps consume about one watt per candle, which is much better than the old carbon lamps whose "efficiency" was not better than one third of a candle per watt. In gas-filled lamps it is possible to heat the filament nearly to 3000°C , and the consumption is then below 0.7 watt per candle.

641. The mercury vapor lamp. An electric discharge through mercury vapor may produce luminescence at the low temperature of 140°C . This is a relatively cold light and highly efficient. If the current is increased, the temperature rises, and a much brighter though less efficient light is obtained. The color of this light appears greenish to some eyes and violet to others, because its visible light is chiefly in strong green and violet spectral lines.

One type of mercury lamp is made by exhausting a tube of glass containing a very small quantity of liquid mercury which fills the tube with vapor. Electrodes are sealed into the ends, and the vapor is made luminous by the passage of an alternating current. This current is applied by a high-tension transformer whose voltage depends upon the length of the tube. Such lamps are very efficient and are used for display purposes.

Another type of mercury lamp depends upon the formation of a true arc. These lamps have two pools of mercury in an exhausted bulb, with platinum wires sealed into the glass to make contact with the mercury. Fig. 60 illustrates a convenient laboratory model. It must be connected to a direct current supply and requires a

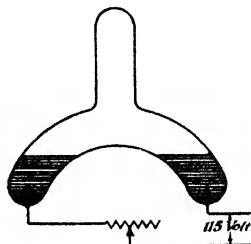


Fig. 60.

potential of about 25 volts between terminals. The current is controlled by a rheostat (variable resistance) when it is operated on a commercial 115 volt circuit. With such low voltages the lamp is not self-starting, and the current is established by tilting the tube until the two pools of mercury unite momentarily. Then as they separate, the arc is struck and continues with one pool acting as anode and the other as cathode.

The light of the mercury arc is so different from daylight that its use is limited to purposes for which it is not necessary to see colors with their normal values. It is used in some factories because of its efficiency, and in some photographic studios because its violet component is highly actinic and greatly shortens the necessary exposure.

In addition to its visible light, the mercury arc is peculiarly rich in ultraviolet rays, but these are almost completely stopped by the glass of the tube. To permit the emission of this light, such lamps are often made of fused quartz, which is very transparent to the ultraviolet. In this case the emergent light is a powerful ionizer of gases, causes many substances to fluoresce brilliantly, and is valuable in the treatment of certain skin diseases. But it is very dangerous to the eyes, which should be protected by spectacles of ordinary glass.

642. The neon lamp. The gas neon at low pressure is even more easily excited to luminescence by an electrical discharge than mercury vapor. A slender tube four feet long requires only 2000 volts, alternating, and consumes very little energy. The color is a warm reddish orange, and is much more pleasing than that of the mercury arc. Neon-filled tubes, bent to form letters and other patterns, are much used for advertising purposes and display. Other inert gases are also used for display and give other colors—blue, green, and so on.

A new form of neon tube, used as a "night lamp" because of its small consumption of energy, screws into an ordinary socket and is designed for commercial alternating voltages. Within the nearly spherical bulb, which contains neon at low pressure, is a metal cylinder cut in two by a small longitudinal gap. Each half cylinder is connected to one terminal, and a negative glow surrounds them alternately when the voltage approaches its peak value, while a direct voltage forms a glow over only one of the half cylinders. Thus an alternating e.m.f. produces a pulsating flickering of twice the commercial frequency, which is much too fast for the eye to detect under ordinary conditions.

643. The sodium vapor lamp. A lamp using sodium vapor is now manufactured in two forms, one for direct, the other for alternating

currents. It delivers 6500 lumens with an expenditure of only 100 watts, a remarkable efficiency which corresponds to five candles per watt. Its light has a pleasant color and affords a very high degree of visibility. In starting the lamp, a discharge passes through a trace of an inert gas contained in the tube, and the heat thus developed vaporizes some of the metallic sodium, which thereafter serves as the carrier of the current. Sodium vapor is also used in connection with the mercury arc, to enrich its spectrum and yield a light more resembling daylight.

644. Electrolysis. The electric current, as we have seen, can be carried through metals, gases, and liquids. Molten metals conduct like solids, but other liquids, when they carry a current, do so by a process known as electrolysis. This means that the liquid must contain carriers, or **ions**, which are either charged atoms or charged molecules, and are produced by the splitting up of certain chemical compounds dissolved usually in water. The passage of the current by means of these carriers results in chemical reactions which do not occur either in metallic conduction or discharges through gases.

The presence of ions in certain solutions called **electrolytes** is due to a phenomenon known as **dissociation**. This was discovered by the Swedish physical chemist Arrhenius. He found that in certain solutions, such as those of salts or acids in water, the osmotic pressure is slightly higher than it should be when calculated from the known concentration. This discrepancy is more pronounced the greater the dilution, and is due to the tendency of the molecules of the solute to dissociate or separate into two ions, one positively and one negatively charged. Thus the total number of dissolved particles is increased, and the osmotic pressure rises. These ions form a perfectly definite proportion of the whole solute at a given concentration, although they are continually recombining as other molecules split up and take their place. This process produces a statistical equilibrium very much as the percentage of divorced couples in a very large city remains approximately constant, though the particular individuals concerned differ from year to year.

If a solution is made increasingly dilute by adding water, there is an increase in the *percentage of dissociation*. In other words, with higher dilution, the number both of whole molecules and of ions in a given volume decreases, but the *proportion* of ions to whole molecules increases. This is because the tendency to split up is at least as great as before, while the chance of *recombination* grows steadily less owing to the greater average separation of any two ions of opposite sign. So

if we double the volume of water we do not quite halve the osmotic pressure.

The reason why certain molecules tend to split up into ions in water, is probably due to the fact that they are normally held together by electrostatic attraction according to Coulomb's law (equation (4), Article 602). This force varies inversely as the dielectric constant, K , of the medium by which the charges are surrounded. Water has an extremely large dielectric constant; therefore the force between the charges is supposed to be so much weakened that occasionally a molecule breaks down spontaneously into its component ions.

645. Migration of the ions. It is the presence of the charged ions which makes it possible to pass electricity through a solution. The positive ions are atoms or groups of atoms which have lost one or more electrons when the dissociation took place. They are drawn toward the negative plate of the electrolytic cell under the influence of electrostatic attraction. The negative ions are atoms or groups of atoms which have gained one or more electrons, and these move

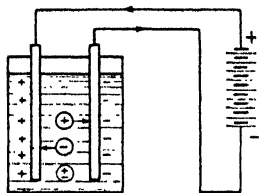


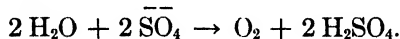
Fig. 61.

toward the positive plate, as represented in Fig. 61. The neutral undissociated molecule, however, does not move under the influence of the field. As an illustration, suppose sulphuric acid is dissolved in water. At once, depending only on the concentration, a certain definite proportion of H_2SO_4

molecules split up into H^+ and SO_4^{--} ions, the former being charged positively, and the latter negatively. There are, moreover, two hydrogen ions to every "sulphion." Each carries an elementary positive charge due to loss of an electron. The sulphion has an excess of two electrons which represent its double unsaturated valence. The passage of these charged particles through the solution under the influence of the field is known as the **migration of the ions**. They move at different speeds of a few centimeters an hour, but both speeds depend upon the voltage impressed upon the cell, the distance between the plates, the concentration, and the temperature.

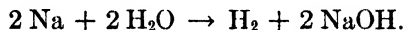
646. Electrolytic reactions. When an ion reaches the plate attracting it, it gives up its charge. This is neutralized by an opposite charge sent there from the battery. Then a fresh unit from the circuit takes its place on the plate, whose charge must be kept constant if the flow is to continue. The discharged ions are either liberated as an un-ionized gas, such as H_2 , or they combine chemically with the solvent

or with the plate itself. If we have dilute sulphuric acid between platinum electrodes, neutral hydrogen gas formed from the discharged $\overset{+}{\text{H}}$ cations is evolved at the negative plate, or **cathode**. At the positive plate, or **anode**, of the same cell, the anions of SO_4 combine with water to form a molecule of oxygen gas and two fresh molecules of H_2SO_4 , according to the reaction:



The neutral oxygen gas is liberated at exactly half the rate of the hydrogen, because, in order to form one molecule of O_2 , it takes two SO_4 ions, each of which was dissociated from two hydrogen ions. Thus we see that the total amount of acid is unaltered, but the water is steadily decomposed.

It should be noted that while in the ionic state, a substance behaves very differently from the way it does in its ordinary form. Sodium, for instance, as an ion, passes through water without reacting with it as ordinary metallic sodium would do. But when it reaches the cathode it gives up its charge and then behaves in the usual way, forming hydrogen according to the reaction:



If this sodium ion were derived from Na_2SO_4 , the SO_4 would act just as in the case of sulphuric acid, and liberate O_2 by reacting with the water at the anode.

When chlorine is the anion from sodium chloride, a different type of reaction occurs. In this case some of the chlorine ions combine with water to form hydrochloric acid and oxygen gas, though most of them are liberated as chlorine gas.

647. Reactions with the electrodes. We have just seen that though the solute produces a variety of ions, the final result may be the same, that is, the liberation of hydrogen and oxygen at the electrodes. But this is not necessarily the case, for if the electrodes are of a substance which can combine chemically with one of the ions, or if the positive cation is a metal which does not combine with water, very different results take place. Copper plates, for instance, in dilute sulphuric acid, ultimately prevent the evolution of either oxygen or hydrogen. The SO_4 ions unite with metallic copper to form CuSO_4 , as is seen by the blue color which rapidly spreads out from the anode, while the $\overset{+}{\text{H}}$ ions are liberated for a time as a gas at the cathode.

But this soon ceases. As more and more copper sulphate is formed, the cation becomes copper instead of hydrogen and is deposited on the cathode as a fresh metallic coating. This process then continues with no change in the strength of the CuSO_4 , once it has been formed from the original acid, until the anode has wasted away, and the cathode has gained a corresponding weight.

648. Electroplating. If the electrodes are of platinum in a solution of CuSO_4 , the cathode becomes coated with copper, while oxygen resulting from $2 \text{H}_2\text{O} + 2 \text{SO}_4$ is liberated at the anode. This continues until all of the solute has been decomposed, leaving only water, and if the current is then reversed, the copper deposit will be re-absorbed and deposited on the other electrode.

If a copper anode is used, the process of electroplating the platinum cathode continues in the same manner as described in Article 647, accompanied by a gradual wasting away of the anode. This then is the usual method of electroplating. The anode is of the metal to be deposited, the cathode the object to receive the metallic coat, and the electrolyte is a salt of the same metal. Gold, silver, nickel, and copper are the substances most commonly deposited in this manner. Gold and silver cyanides are used as the electrolytes in plating with these elements, nickel is deposited from a double sulphate of nickel and ammonium, and copper from copper sulphate. Copper, moreover, is refined by a similar process from an anode of the crude metal to a cathode which builds up what is known in the market as "electrolytic copper."

649. Faraday's laws of electrolysis. In 1833, Michael Faraday first formulated the laws by which quantitative calculations of electrolysis were made possible. They are:

1. *The total mass of the substance liberated at either electrode is proportional to the quantity of electricity which has passed through the cell.*

2. *The total mass of the substances at either pole liberated by the same quantity of electricity is proportional to their chemical equivalents.*

The first law tells us that with a given substance, $M \propto Q$. Hence

$$M/M' = Q/Q', \quad (1)$$

where M and M' are the masses of a given substance deposited by the quantities Q and Q' respectively.

In the second law, chemical equivalent means the atomic weight divided by the valence. It is also called **combining weight**. If then we compare the deposit of two different substances produced by the

same quantity of electricity, the second law may be expressed as

$$M/M' = \frac{w}{v} / \frac{w'}{v'}, \quad (2)$$

where w is the atomic or molecular weight, and v is the valence.

These two laws may be combined in the single expression

$$M = C \left(\frac{w}{v} \right) Q,$$

or

$$M = C \left(\frac{w}{v} \right) It = zIt, \quad (3)$$

where C is a constant of proportionality, and $C \left(\frac{w}{v} \right)$ is the **electrochemical equivalent** of the ions denoted by z . The value of C is usually determined by the experimental plating of silver from a solution of silver nitrate. This process is very dependable, and the result is a deposit of 0.0011180 gram per coulomb. The atomic weight of silver is 107.88, and it has unit valence. Therefore, substituting these values in equation (3), we obtain

$$0.0011180 = C \frac{107.88}{1},$$

or

$$C = 10,363 \times 10^{-9} \text{ gram per coulomb.}$$

Now that C is known, we may calculate the mass of any ion liberated by the passage of a known quantity of electricity. Then let the amount liberated be chosen so as to be as many grams as the numerical value of the combining weight of its substance, w/v . In other words, suppose a *gram equivalent* is to be liberated; then equation (3) becomes

$$\frac{w}{v} = C \left(\frac{w}{v} \right) Q.$$

$$\therefore Q = \frac{1}{C} = F.$$

Taking the value of C just determined, we find that its reciprocal Q equals 96,494 coulombs. This is the quantity of electricity required to liberate a gram equivalent of any ion. It is the natural electrochemical unit of quantity, usually designated by F , and is known as the **faraday**, or "Faraday's constant."

By substituting in equation (3) the proper value of w/v of the cation, Faraday's equation enables us to calculate the weight of a deposit in electroplating from any electrolyte. If the current is expressed in amperes and the time in seconds, M is given in grams. If,

however, the *relative* amounts of two or more substances deposited by the same quantity of electricity are required, the calculation is still simpler, for then we have the continued proportion

$$M_1 : M_2 : M_3 : \dots : M_n : \frac{w_1}{v_1} : \frac{w_2}{v_2} : \frac{w_3}{v_3} : \dots : \frac{w_n}{v_n}.$$

650. Significance of Faraday's constant. As we have shown that the constant C is the reciprocal of F , the electrochemical equivalent, z , of any ion, defined above as Cw/v , may be expressed by

$$z = w/Fv. \quad (1)$$

The product Fv is the charge carried by a gram molecule of any substance, and as there are 6.06×10^{23} molecules in a gram molecule (Avogadro's number, N) we may obtain the charge Q' carried by a single ion, by dividing Fv by N , thus

$$Q' = Fv/N. \quad (2)$$

If the ion is singly ionized, $v = 1$, and denoting this particular charge by e , we obtain

$$e = F/N. \quad (3)$$

Taking the accepted values of F and N given above, we find that $e = 1.59 \times 10^{-19}$ coulombs, or 1.59×10^{-20} emu. This is the basic charge of electrolytic conduction. Doubly ionized atoms or molecules carry two such charges, trebly ionized atoms or molecules carry three, and so on.

The ratio of the charge of an ion to its mass is an important quantity in electron theory. This is easily calculated as follows: The mass m of an ion is always given by w/N (see Article 203), and its charge is given by (2) above. Then

$$\frac{Q'}{m} = \frac{Fv/N}{w/N} = \frac{Fv}{w} = \frac{1}{z}. \quad (4)$$

Thus the ratio of the charge to the mass of an ion equals the reciprocal of its electrochemical equivalent. If the ion is singly ionized, the ratio e/m equals $1/z$, and in the case of hydrogen, when $w = 1$ and m_H is its mass,

$$e/m_H = 1/z = F. \quad (5)$$

651. The international ampere. For commercial purposes the ampere has been defined by law as *the unvarying electric current which, when passed through a solution of nitrate of silver in water, deposits silver at the rate of 0.001118 gram per second.* The cell in which this is accomplished is carefully described, for the deposit

varies slightly with its construction. This unit of current is not to be thought of as anything but a convenience, and in no way replacing the true definition upon which it is based. The scientific unit already defined, however, is very difficult to measure with precision, and manufacturers find the international ampere a valuable substitute. It agrees closely enough with one tenth of the absolute unit of current for most purposes.

652. Polarization. There are several secondary phenomena connected with electrolysis which are grouped under the general name of **polarization**. The most important of these is an opposing e.m.f. set up within the cell. Suppose a current from a battery, giving a constant e.m.f. of several volts, is sent through an electrolytic cell containing dilute sulphuric acid between platinum electrodes. If the current is measured by a suitable meter, it will be found that on closing the circuit it decreases rapidly from its initial strength, and then tends to remain constant at a much smaller value. This is caused by the accumulation of the ions at the two electrodes, which thus become equivalent to plates of oxygen and hydrogen. They then behave like a battery producing a counter-e.m.f. which acts against the impressed e.m.f. to reduce the flow of the current. The value of the counter-e.m.f. in the case supposed is 1.119 volts, and the impressed voltage must exceed this value in order to maintain the current after the polarization has been established. If Ohm's law is to be applied to such a circuit, it must be modified to read

$$I = \frac{E - E'}{R},$$

where E is the impressed e.m.f., E' the counter-e.m.f., and R the resistance of the cell.

To make the current flow against the counter-e.m.f. of polarization in an electrolytic cell, energy has to be supplied. This energy is needed to bring about the chemical changes that occur. If no energy were needed to decompose water, we might violate the conservation of energy in the following way: Set up a large cell with platinum electrodes very close together and pour in acidulated water. Then send a current through the cell. As the cell is supposed large and the plates are close together, the resistance is very low and very little power would appear to be necessary. Then the oxygen and hydrogen given off might be used as fuel giving much more energy than was needed to produce them. This is contrary to the law of the conservation of energy, and is impossible. The explanation is that extra work

is done in sending the current against the counter-e.m.f. of polarization, and this extra work is used in breaking up the water molecules. So the fuel value of the produced gases was not obtained for nothing after all!

653. Conductivity of electrolytes. If a current is sent through a nonpolarizable electrolytic cell, we may apply Ohm's law and calculate the resistance from the current and from the observed potential drop across its terminals. The reciprocal of the resistance is the conductance of the cell. If this is measured, and if the area of the plates and their distance apart are known, the conductivity k may be obtained from the relation $1/R = kA/l$.

In comparing the conductivity of different electrolytes, it is necessary to use corresponding concentrations. This is accomplished by taking the same number of gram molecules per unit volume. If we dissolve a gram molecule, which is a mass in grams numerically equal to the molecular weight, in a liter of water, we obtain a **molar solution**. The number of molecules present is then always equal to Avogadro's number, as it is the number of molecules in a gram molecule of any substance.

The number of gram molecules per liter, or **moles per liter**, is called **molecular concentration**. If one mole is dissolved in a liter of water, the solution is called **normal**. Other concentrations are called "half normal," "tenth normal," and so forth. The number of moles per cubic centimeter is one thousandth of the molar concentration.

654. Molecular and equivalent conductivities. Molecular conductivity is defined as the ordinary conductivity divided by the molar concentration per cubic centimeter; therefore

$$k_m = k \div \frac{m}{1000} = \frac{1000k}{m}, \quad (1)$$

where k_m is the molecular conductivity, k the ordinary conductivity, and m the molecular concentration in gram molecules per liter.

Equivalent conductivity is defined as the ordinary conductivity divided by the concentration in gram equivalents per cubic centimeter, or $m/1000v$. Therefore equivalent conductivity equals the molecular conductivity times the valence.

655. Effect of dilution on conductivity. If we measure the resistance of a given electrolyte between plates at a fixed distance apart, it would seem natural to expect that both the molecular and equivalent conductivities would remain constant with increasing dilution. One would expect the conductivity to decrease at the same rate

as the concentration m , leaving k_m the same. But this is not the case. The molecular conductivity increases, and may become several times as great as its original value, as we approach infinite dilution. This is because, although the number of molecules is proportional to m , the number of ions is not, but becomes relatively greater as the average distance between them increases.

SUPPLEMENTARY READING

A. W. Hirst, *Electricity and Magnetism* (Chap. 7), Prentice-Hall, 1937.

Maass and Steacie, *An Introduction to the Principles of Physical Chemistry* (Chap. 13), Wiley, 1931.

Creighton and Fink, *Electro-Chemistry*, Volume I, Wiley, 1924.

PROBLEMS

1. Calculate the resistance of 3, 6, 8, and 9 ohms in parallel. *Ans.* 1.36 ohms, nearly.

2. What is the total current if 106 volts are applied to the circuit of Problem 1, and what is the current in the 8-ohm branch? *Ans.* 78 amperes; 13.25 amperes.

3. What is the current when 50 volts are applied to 1 ohm in series with a branched circuit of 2, 3, and 4 ohms in parallel? *Ans.* 26 amperes.

4. In Problem 3 calculate the fall of potential across the three parallel resistances and the current in each branch. *Ans.* 24 volts; 12, 8, and 6 amperes.

5. The leads (#9 B and S gauge) connecting a generator with a set of fifty 200-ohm lamps in parallel are each 250 ft. long and have a resistance of 8×10^{-4} ohm per foot. What are the current and drop on the line with a generator e.m.f. of 110 volts? *Ans.* 25 amperes; 10 volts.

6. The average temperature coefficient of the resistance of copper is 0.00428. Calculate the current in Problem 5 if the temperature of the leads is increased by 30° C. *Ans.* 24.7 amperes.

7. At 8 cents per kilowatt-hour, how much do 2 million coulombs cost when the e.m.f. is 110 volts? *Ans.* \$4.89.

8. What is the cost of 2 million calories at 8 cents per kilowatt hour? *Ans.* 18.6 cents.

9. How long must a current of 3 amperes flow through an electrolytic cell to decompose 2 g of water? *Ans.* 1 hr. 59 min. 8 sec.

10. In a copper voltameter (a cell which measures coulombs by the weight of metal deposited) 0.593 g of copper (valence = 2) is deposited by a current of 2 amperes in 15 min. Calculate the atomic weight of copper. *Ans.* 63.58.

11. If 0.3 g of hydrogen is liberated by a certain current in 10 min., how long will it take the same current to deposit 15 g of gold in a gold plating bath? (The atomic weight of gold is 197.2, its valence is 3.) *Ans.* 7 min. 36 sec.

CHAPTER 48

Batteries

656. Electromotive force. The fall of potential when a current flows through a wire is given by the product of the current and the resistance, as was explained in Article 625. Therefore IR vanishes if there is no current, or if there is a perfect short circuit, when $R = 0$. This quantity is in the nature of a reaction caused by the resistance. In fact, it is sometimes called **resistance reaction**.

In order that there may be a reaction, there must be an action. The "action" is called the **electromotive force** which, as has been explained, is the difference of potential due to some agent such as a battery. It is measured, of course, in the same units as the reaction, that is, in volts, and is not a "force" in spite of its name. Sources of electromotive force are like pumps in which the flow is from low to high levels, while the resistance reaction is like the fall of level in a stream flowing under the influence of gravity.

657. Internal resistance. In any source of electromotive force there must be some resistance. This *internal resistance* has no effect on the e.m.f. produced, but does alter the difference of potential measured across the terminals of the source when the current is flowing. The internal drop caused by this resistance r is equal to Ir ,

and is a tax upon the actual e.m.f. developed. So the terminal difference of potential is given by

$$\Delta V = E - Ir,$$

where E is the electromotive force of the source on open circuit, and ΔV is the difference of potential of its terminals when a current I is

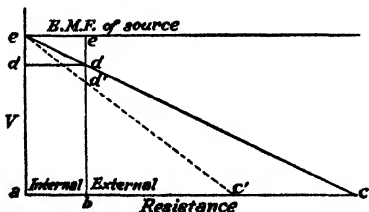


Fig. 62.

flowing through it. This may be represented graphically as in Fig. 62, where potential is plotted as a function of resistance. As the current is the same throughout the circuit, the line edc , which represents the potential at any point, is straight. It slopes from the point representing the e.m.f. of the source at e , down to c , which must be thought of

as joined again to a , where the source again raises the level to e . The internal drop of potential is ed , and the external is db , which also equals the terminal voltage on closed circuit. If the external resistance is smaller, the current is stronger and the line is steeper. Then c is closer to a , as at c' , and the internal drop ed' is increased, while the terminal potential difference $d'b$ is decreased.

The fall of potential within a cell when it is delivering a current may involve a serious loss of energy if its resistance is large. Also, more cells are required to send a given current through a given external resistance. Therefore both primary and secondary (storage) cells are designed to have as low an internal resistance as possible. A dry cell is commonly tested for its internal resistance when it is sold, because this is an index of its freshness—the lower the resistance the fresher the cell. This test consists in short-circuiting the cell through an ammeter. Thus if the ammeter reads 25 amperes, and if the e.m.f. is 1.5 volt, the internal resistance is $1.5 \div 25 = 0.06$ ohm, which is a fair showing, though some cells give 30 amperes on short circuit.

If two or more cells are connected in series, their total resistance and total e.m.f. are the sums of all the individual resistances and voltages. Therefore four cells, like the one supposed above, when connected in series, would have an e.m.f. of 6 volts and an internal resistance of 0.24 ohm, but the short-circuit current would still be 25 amperes, as with one such cell. If, however, n cells are connected in parallel, the total e.m.f. is that of one cell only, but the internal resistance is one n th of that of each, and the short-circuit current is n times as large.

In order to calculate the current through an external resistance R , the total e.m.f. is divided by R plus the internal resistance of the whole battery of cells. If they are in series, $I = nE/(R + nr)$, where r is the internal resistance of each cell. If they are in parallel, $I = E/(R + r/n)$. If there are m strings of cells with n cells in series in each string, and the strings are in parallel, then the internal resistance of this series-parallel arrangement is clearly nr/m , and the total current is given by

$$I = \frac{nE}{R + \frac{nr}{m}},$$

or

$$I = \frac{mnE}{mR + nr}. \quad (1)$$

Thus if a total of mn cells is available, the current may be varied according to how they are grouped. Obviously I is a maximum when the denominator in equation (1) is a minimum, since the numerator is constant at constant voltage. This occurs when $mR = nr$, because the product of the two factors (mn) and (Rr) is constant, and in general the sum of two factors whose product is constant is least when they are equal. Therefore the current is maximum when $R = nr/m$, or when the external and internal resistances are the same. If, then, R and r are known, we may group the nm cells so as approximately to fulfill the theoretical requirement.

658. The "galvanic" couple. The earliest known source of electromotive force capable of maintaining a continuous flow of electricity was accidentally discovered in 1786 by Galvani, a professor of anatomy in the University of Bologna. He observed that freshly prepared frogs' legs near an electrical machine twitched when the discharge took place. He also found that if they were hung by a copper hook passing through the lumbar nerve, the muscle contracted violently when the leg came in contact with the iron rail of his balcony. This second discovery, though resulting from the first, was really accidental also, but it led to an investigation of the so-called "contact" electromotive force.

Galvani had rightly concluded that the twitching of the frog's leg was due to an electric stimulus, and assumed that the current originated in the animal tissues as a sort of fluid.

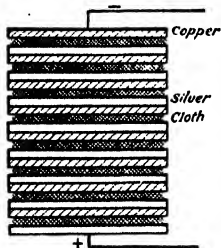


Fig. 63.

But Volta, of the University of Padua, disputed this conclusion, and showed that the source of electrification was due to the contact of the two dissimilar metals, copper and iron, acting through the moist leg of the frog. In his **voltaic pile** he amplified this effect by having a series of such contacts, or *couples*, arranged as shown in Fig. 63. Here pieces of cloth moistened with dilute acid or a solution of salt, or even plain

water, separate every alternate pair of silver and copper discs piled on top of each other. The moistened cloth takes the place of the frog's leg, and the e.m.f. is directed from copper toward silver across the pad. A fairly large voltage may thus be built up by using a large number of couples to form the voltaic pile.

Volta thought that the moistened pad acted merely as a conductor, and that the e.m.f. was due chiefly to the junction of dissimilar metals. But it is now very certain that although this effect exists, it is

extremely small. The real source of the e.m.f. was shown by Faraday and others to be the chemical action caused by the moistened pad.

When dissimilar metals are separated by an electrolyte or a very thin layer of a gas which acts upon them, they develop a potential difference. Various metals may then be arranged in a series in which those coming first are electropositive to those which follow. If the metals are brought together in air, the series is: zinc, lead, tin, nickel, iron, copper, gold, silver, graphite. Thus tin is electropositive to copper in air, but electronegative to zinc. In an atmosphere of hydrogen sulphide the series is quite different, with zinc following copper, among other changes. Therefore the order of the series depends upon the nature of the medium surrounding the metals, and even upon its concentration. In dilute nitric acid, zinc is electropositive toward tin, but it is electronegative in strong acid.

659. The voltaic cell. If instead of separating two metal plates by a moistened pad, they are plunged into a jar containing an electrolyte such as dilute sulphuric acid, the same difference of potential appears between them, as in the voltaic pile. Taking copper and zinc, for instance, and connecting them externally by a wire as in Fig. 64, we find that a current flows in the wire from copper to zinc, and from zinc to copper through the dilute acid. An accepted elementary explanation of this action is due to Nernst, and accounts for the production of an e.m.f. and consequent flow of current when two dissimilar metals are placed in an electrolytic solution. The chief features of Nernst's theory are given in the following article.

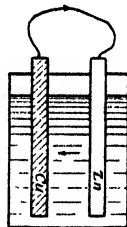


Fig. 64.

660. Solution pressure. When a metallic body is plunged into pure water or an electrolyte, it tends to throw off ions into the liquid with what is known as **solution pressure**. Since the metallic ions are all charged positively, this loss of positive electricity by the body gives it a negative charge, while the liquid near the plate acquires a positive charge. These charges go on increasing until equilibrium is established by the electrostatic attraction between the layer of free ions which forms in the liquid, and the equal and opposite charge upon the metal plate. These charges form a "double layer," like the plates of a charged condenser, as shown in Fig. 65, and establish a field directed toward the plate. As the charges of the layers increase, the field strength increases until it reaches the equilibrium value. Then no more positive ions are able to move out against it. In the case of pure water, there is no other effect, and a definite potential

difference between the layers is quickly established, depending only upon the solution pressure of the particular metal used. But if the liquid is an electrolyte, there is also an osmotic pressure exerted by the positive ion of the solute, which opposes the solution pressure. If



Fig. 65.

this is greater than the solution pressure, the solution ions "condense" (as a vapor condenses) on the metal, giving it a positive charge, and leave the liquid negatively charged, thus forming a double layer with the field directed from metal to liquid. If, however, the solution pressure predominates, the metallic ions escape from the plate into the liquid, until a balance is established between solution pressure P_s on the one hand, and osmotic pressure P_o added to electrostatic attraction F on the other. This may be written symbolically thus:

$$P_s \rightarrow P_o + F,$$

where P_s and P_o are constant, and F varies from zero to a maximum value F_m when equilibrium is established. Then

$$P_s = P_o + F_m.$$

Of course, P_o differs with the nature of the solution, and the same metal may become either positively or negatively charged in different electrolytes. The most familiar case occurs when the positive ions of the solution are the same as those from the metal, as when zinc is plunged into zinc sulphate, and copper into copper sulphate. In one case the zinc plate is negative to the liquid, because $P_s > P_o$. Each escaping zinc ion leaves two electrons on the plate, while the neighboring liquid acquires a positive charge. In the other case, copper is positive to the liquid, because $P_o > P_s$, so $P_o = P_s + F_m$. The copper ions in the solution tend to collect on the plate, from which they take away two electrons each. This charges it positively, while the liquid near the plate becomes negatively charged with an excess of SO_4^{--} ions.

661. The Daniell cell. In this cell a copper rod or plate is immersed in a solution of copper sulphate contained in a porous cup, and surrounded by a solution of zinc sulphate in which a cylinder of sheet zinc surrounds the cup. This is equivalent to the arrangement shown in Fig. 66, where the porous cup is replaced by a plane porous partition. This separates the liquids but allows free passage of the ions.

On open circuit, equilibrium is soon established. The more electropositive metal, zinc, by virtue of its higher solution pressure,

acquires a negative charge, and the copper, whose solution pressure in copper sulphate is very low, acquires a positive charge. This means that although zinc is electropositive to copper in the voltaic series in air, in a Daniell cell it acquires a negative charge, and its potential falls below that of the liquid, while that of copper rises above it.

If now the circuit is closed, the charges on the plates neutralize each other, and the attracted charges in the liquid are freed from the electrostatic force which held them near the plates. At once new zinc ions escape into the zinc sulphate solution, and new copper ions are deposited from the copper sulphate solution upon the copper plate. The plates are thus charged again only to be discharged, and a continuous current around the circuit ensues. In the liquid the positive zinc ions, continually liberated, carry their charge toward the porous partition. There they unite with the negative sulphions liberated when the copper ions go out of solution to form metallic copper. This union results in the formation of neutral zinc sulphate molecules, and the whole process is equivalent to a conduction current through the liquid, having the same magnitude as that in the wire.

As a result of this process the zinc sulphate solution grows more concentrated at the expense of the zinc plate, while the copper sulphate solution grows steadily weaker. To offset these losses copper sulphate crystals must be present in the porous cup, so as to maintain the solution concentrated, and the zinc plate which wastes away must ultimately be replaced.

As osmotic pressure tends to weaken the negative charge on the zinc plate and strengthen the positive charge on the copper plate, it

Metal	Potential
Magnesium....	- 1.45
Zinc.....	- 0.522
Cadmium.....	- 0.152
Lead.....	+ 0.126
Copper.....	+ 0.581
Mercury.....	+ 0.990
Silver.....	+ 1.024

is clear that the potential of the cell is a function of the concentrations of the two electrolytes, and should decrease as the zinc sulphate solution grows stronger, and the copper sulphate weaker. That this is the case is an important confirmation of Nernst's theory.

Many cells similar to Daniell's, but using different metals and electrolytes, are possible. The accompanying table gives values by which the voltages of such cells at 18° C may be computed. It gives the

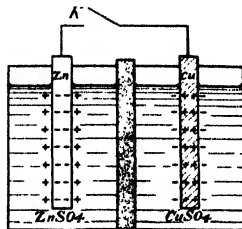


Fig. 66.

difference of potential in volts between a metal and a solution of its sulphate, all having the same molecular concentration.

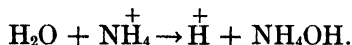
662. Polarization of a cell. If dilute sulphuric acid is used as the electrolyte in a zinc-copper cell, it amounts to a couple in the voltaic pile. The process in this case is somewhat different and much less satisfactory than with the Daniell cell.

The zinc and copper plates acquire negative and positive charges as before, but now, with closed circuit, positive hydrogen ions form on the copper plate, instead of the copper ions as in the Daniell cell. This sets up an opposing electrostatic field. It tends to neutralize the field that develops on open circuit, and the external current falls rapidly to a very small value. Polarization of this kind is explained by Nernst's theory, for hydrogen has a very large solution pressure, resists being thrown out of solution, and tends to come back again in opposition to the field between copper and the acid. In fact, we are forming what amounts to a zinc-hydrogen couple in which hydrogen is more electropositive than zinc, and therefore tends to drive a current through the cell from hydrogen to zinc.

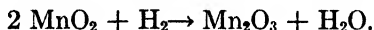
To eliminate this difficulty, all effective primary cells use some depolarizing agent which absorbs chemically the discharged cations at the positive pole of the battery as fast as they are formed. In the cell we are discussing, copper sulphate would act in this way. If it surrounds the copper plate, the hydrogen ions unite with the sulphion of CuSO_4 to form sulphuric acid and copper, thus: $\text{H}_2 + \text{CuSO}_4 = \text{H}_2\text{SO}_4 + \text{Cu}$. The liberated copper is deposited on the copper plate, and the sulphuric acid goes to restore the concentration of the electrolyte. Thus modified, such a cell is rapidly transformed into a Daniell cell, provided a porous membrane is used. Then the zinc ions form zinc sulphate with the sulphions of the solution, while the loss of an acid molecule is exactly compensated by its formation from the copper sulphate, as noted above.

663. The Leclanché cell. Here the electrodes are zinc and carbon instead of zinc and copper, and thus yield a higher voltage than the Daniell cell. This might be expected from the voltaic series given in Article 658, where graphite (a form of carbon) is seen to be farther from zinc in the "contact" potential series than is copper. The electrolyte is a solution of ammonium chloride, whose ions are NH_4^+ and Cl^- . The chlorine, or negative, ion moves toward the zinc electrode and, when the cell is delivering a current, combines with the zinc to form zinc chloride. The ammonium ion tends to react with

water at the carbon plate to form ammonium hydroxide and give up one atom of hydrogen according to the reaction



This would cause a counter-e.m.f. of polarization, as in the case of the Daniell cell, and to prevent it, the carbon is surrounded with manganese dioxide as a depolarizer. This oxide is reduced by the liberated hydrogen to form manganic oxide and water, thus:



The so-called "dry cell" is not dry, but is essentially a Leclanché cell with the electrolyte, composed of several components in addition to ammonium chloride, made up as a moist paste or jelly. When it becomes really dry, the cell ceases to function.

664. The standard cadmium cell. This cell, invented by Edward Weston of New Jersey, was adopted in 1908 by the International Electrical Congress as a standard of e.m.f. At 15° C it gives 1.0185 volts at its terminals on open circuit, or when no current is flowing, so that the volt may be defined as 1/1.0185 of the e.m.f. of a standard Weston cell. This "legal volt," like the legal ohm and ampere, is very convenient for commercial and laboratory practice. However, it is not based on fundamental physical concepts and differs from the true volt, which is 10⁸ times the absolute unit already defined in Article 626.

The Weston cell is made according to exact specifications, as shown in Fig. 67. The positive electrode is mercury covered by a paste of mercurous sulphate on which are loose crystals of cadmium sulphate. The negative electrode is an amalgam of cadmium and mercury, also covered with cadmium sulphate crystals. The electrolyte is a concentrated solution of cadmium sulphate. This combination is tightly sealed in an H-tube, as indicated, with platinum wires let in through the glass to make contacts with the electrodes. When properly constructed, this cell will give its rated voltage for years, provided no appreciable current is drawn from it, and as it has an extremely low temperature correction, its e.m.f. may be regarded as independent of ordinary temperature variations in the laboratory.

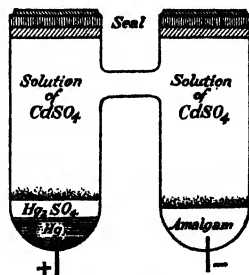


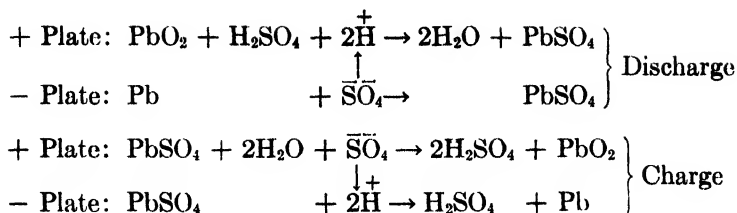
Fig. 67.

665. Storage cells. If an external electromotive force is opposed to that of a Daniell cell, a current will flow through it in opposition to the cell's e.m.f., provided the applied voltage is the higher one. When this occurs, the zinc ions of the zinc sulphate solution are deposited on the zinc plate against its solution pressure, and this electrode gains in weight, while copper ions are forced into solution from the copper plate against their osmotic pressure in the copper sulphate solution. The removal of the zinc ions from the solution of zinc sulphate involves a progressive liberation of sulphions which unite with the copper ions, liberated from the copper electrode, to form copper sulphate, thus strengthening the latter while the zinc sulphate solution is being weakened.

It is evident that a Daniell cell which has been in operation for some time, with consequent consumption of metallic zinc and copper sulphate, and with a gain in metallic copper and zinc sulphate, can be restored to its original condition by the reversal of the current. This is the basic principle of the storage cell, which in no sense stores up electricity, but after use can be brought back to its original elementary status by a reversal of the current, when the so-called "charging" process takes place. Of course, it is no more electrically *charged* afterwards than before, but chemical reactions have taken place which bring back its original constitution, and consequently its original electromotive force.

666. The lead accumulator. By far the most successful storage battery yet devised is the "lead" battery. It excels all others in its high voltage, low internal resistance, and in the efficiency with which the electrical energy used in charging can be recovered on the discharge. Its most serious drawbacks are its weight and the fact that if not kept properly charged, it deteriorates rapidly to a point where it is practically impossible to restore it.

This cell is made of two lead grids into one of which a paste of lead peroxide, PbO_2 , is pressed, while into the other are inserted strips of spongy metallic lead. The electrolyte is dilute sulphuric acid which reacts with the electrodes as follows: On *discharge*, the lead peroxide of the positive plate unites with the hydrogen ions and sulphuric acid to form lead sulphate and water, while the metallic lead of the negative plate combines with the sulphions to form lead sulphate also. On *charge*, the reverse process takes place, and the lead sulphate of the positive plate reacts with the sulphions to re-form lead peroxide, while the hydrogen ions reduce the sulphate at the negative plate back into metallic lead. This may be shown symbolically thus:



The vertical arrows indicate the direction of the current, and therefore the direction of migration of the positive hydrogen ions. It will be seen that during discharge the acid in the solution is broken up, and its concentration diminished, while it is regenerated during charge. This useful fact makes it possible to test the condition of a battery by using a hydrometer, which gives the density of the solution. It should be about 1.250 when the cell is fully charged, and should never fall below 1.150 as a result of too heavy a discharge. During the process of charging, a potential of about 2.5 volts is necessary to send a suitable current through the cell in opposition to its back e.m.f. (a trifle over two volts) and to supply the additional energy which appears as heat. When charged, the cell has an e.m.f. also slightly over two volts, but on discharge it falls quite rapidly to 1.9 volts, and then more slowly to 1.8. It should then be charged again before its voltage goes below this value.

As there is a steady formation of lead sulphate on both plates during discharge, and as this compound is highly insoluble and has a very high resistance, it is evident that too great a discharge is fatal to such a battery. The plates then become wholly coated with the white deposit, and a charging current cannot flow through it to bring it back to its original condition.

SUPPLEMENTARY READING

- Wm. H. Timbie, *Elements of Electricity* (Chap. 10), Wiley, 1925.
 A. W. Hirst, *Electricity and Magnetism* (Chap. 8), Prentice-Hall, 1937.
 G. W. Vinal, *Storage Batteries*, Wiley, 1930.

PROBLEMS

1. Six dry cells in series deliver current to a circuit whose resistance is 3 ohms. The e.m.f. of each cell is 1.5 volt and its internal resistance is 0.06 ohm. What are the current and the terminal volts of the battery?
Ans. 2.68 amperes; 8.04 volts.
2. If the cells in Problem 1 are in parallel, what are the current and terminal volts? *Ans.* 0.498 ampere; 1.494 volt.

3. A cell whose e.m.f. is 2 volts is connected in opposition to one giving 1.55 volt. The internal resistance of the former is 0.1 ohm, and of the latter, 0.05 ohm. What are the current and terminal volts? *Ans.* 3 amperes; 1.7 volt.

4. Two cells whose e.m.f. are 2 and 1.5 volts, and internal resistances 0.1 and 0.075 ohm, respectively, are connected so as to send currents in the same sense around their common circuit. What are the current and terminal volts? *Ans.* 20 amperes; zero.

5. Three strings of 4 cells each are connected in series-parallel. The e.m.f. of each cell is 1.5 volts and its internal resistance 0.1 ohm. What current flows through an external resistance of 1.8 ohm? *Ans.* 3.10 amperes.

6. Show how to arrange a group of 96 cells whose internal resistances are 0.1 ohm each, so as to send a maximum current through an external resistance of 0.6 ohm. *Ans.* 4 strings of 24 cells each.

CHAPTER 49

Thermoelectricity

667. The Peltier effect. If a current is sent across the junction of two bars of dissimilar metals, in addition to the Joule heat developed throughout, heat is either evolved or absorbed at the junction. This fact was discovered in 1834 by Peltier, a French physicist (1785–1845). The amount of the Peltier heat in joules is given by $W = E_p It$, where E_p is an electromotive force directed across the contact. If the current flows against E_p , it does work and develops heat which warms the junction. If it flows with E_p , work is done on the current, heat is absorbed, and the junction is cooled. This is a true contact electromotive force, and has a much smaller value than those observed by Volta when chemical action took place.

If the junction is between a copper and iron bar, and if the current flows from iron to copper across the junction, heat is evolved. If it flows from copper to iron, the junction is cooled. Then the Peltier e.m.f. must be directed from copper to iron. This means that the copper must be at a lower potential than the iron, because an electromotive force is always directed from low potential to high. The lower potential of copper is in accord with the usual voltaic series, where both in air and in dilute sulphuric acid, iron is electropositive to copper.

The value of E_p in most couples is of the order of 0.001 volt, and the Peltier heat developed by a current of an ampere is the same as the Joule heat when one ampere flows through a resistance of 0.001 ohm. This would be very hard to observe, because it would be masked by the Joule heat. But if the current is reduced to a milliampere, the Peltier heat is the same as the Joule heat developed by a milliampere in one ohm. Then if the junction has a fairly large section whose resistance is a small fraction of an ohm, the Joule heat is negligible in comparison with the Peltier effect.

The thermocouple which has the largest Peltier e.m.f. is one made of copper and an alloy of bismuth and antimony. It develops 0.022 volt at 25° C.

668. The Seebeck effect. Heating or cooling a junction by an electric current is the inverse of a phenomenon discovered in 1821 by

Seebeck, a German physicist (1770–1831). If two metal bars or wires are connected as in Fig. 68, and if one of the two junctions is heated, a current flows around the circuit. When the metals are copper and iron, the Peltier e.m.f. is directed from copper toward iron.

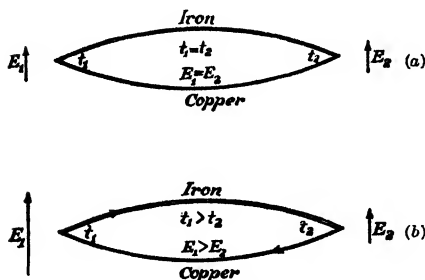


Fig. 68.

If both junctions are at the same temperature, as indicated in (a), the values E_1 and E_2 of E_p are equal and no current flows. But the *Peltier electromotive force rises with the temperature*, so that if the first junction is heated, E_1 becomes larger than E_2 , and a current flows *from copper to iron across the heated junction*.

This current must tend to cool the junction, in accordance with the Peltier effect. If it heated the junction, the rise of temperature would make the current stronger. Then there would be a further rise of temperature and the current would increase of itself indefinitely, which is impossible, as stated by the second law of thermodynamics (Article 263).

669. Thermo-e.m.f. and temperature relations. The current produced by heating one of the junctions, shown in Fig. 68 (b), increases as that junction's temperature rises, reaches a maximum, then decreases to zero, and finally flows in the opposite direction. This current is caused by the thermo-electromotive force, which may be measured by a suitable voltmeter. A curve plotted between this e.m.f. and the temperature is given in Fig. 69. It is symmetrical with respect to a line drawn perpendicular to the temperature axis and passing through its highest point. Therefore the temperature of the heated junction, at reversal, is as far above its value at the peak of the curve as the cooler junction is below it. The temperature of the peak is known as the **neutral temperature**, and that at reversal as the **temperature of inversion**. The neutral temperature is inde-

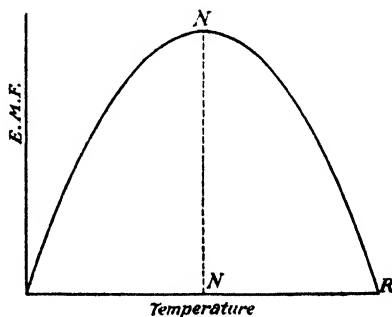


Fig. 69.

pendent of the temperature of the cooler junction. If the cooler junction is taken at a *higher* value, for instance, the thermo-e.m.f. does not reach as high a maximum and reverses sooner. The resulting curve is still symmetrical about the same neutral temperature axis, as shown in Fig. 70. The various curves obtained by keeping the cooler junction at various constant temperatures are all parabolas. In fact, they are arcs of the same parabola that would be obtained by cutting off various amounts by the temperature axis if the first parabola were moved vertically downwards.

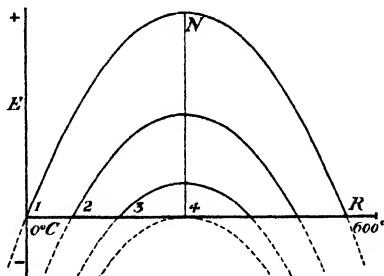


Fig. 70.

vertically downwards. In the extreme case in which one junction is kept at the neutral temperature, the current will flow in the reverse direction (from iron to copper) across the other, whether it is either warmed or cooled, as is evident from the fourth curve.

The fixed neutral temperature of a copper-iron couple is commonly said to be 275°C , though the kind of copper and iron used makes a good deal of difference. The variable inversion temperature is twice this value, or 550° , when the cooler junction is at 0° , but correspondingly higher if it is below 0° , and lower if it is above 0° . Thus, if the cooler junction is at -10° , the heated junction must reach 560° before the current changes direction.

670. The Thomson effect. In 1851, Sir William Thomson (later Lord Kelvin) showed by thermodynamic reasoning that the reversal of the thermo-e.m.f. must be due to a field in opposition to the Peltier e.m.f. This predicted effect was then verified experimentally and named after its discoverer.

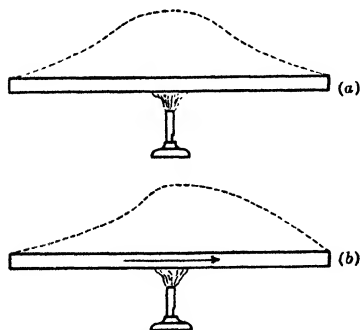


Fig. 71.

The Thomson effect is caused, not by the contact of dissimilar metal bars, but by the unequal temperature of two parts of the same bar. It may be demonstrated as follows: If a copper rod is heated at its center, the temperature rises to a maximum there, and falls off symmetrically on either side, as indicated by the dotted line in Fig. 71 (a).

But if at the same time a current is sent through the bar, the temperature curve is no longer symmetrical, but indicates more heat on the side toward which the current is flowing and less on the other side, as shown in Fig. 71 (b). This is readily explained by assuming a higher potential where the bar is heated, so that the current in flowing up the grade requires work done upon it and so absorbs heat, while in flowing from higher to lower potential it does work and evolves heat. In the case of iron, the opposite effect occurs, and the temperature curve shows a decreased heating on the side toward which the current is flowing. This indicates a lower potential at the heated portion of the bar, and a consequent absorption of heat where the current flows from the low potential of the heated center toward the higher potential of the colder end. There is also an evolution of heat where the current flows toward the heated center.

The potential difference caused by heating some portion of a conductor does not represent an electromotive force as we have used that term. A true e.m.f. tends to cause a current to flow from low to high potential within the region where it is developed. But the potential difference of the Thomson effect does not act in this way. The *tendency* is for the charges created by unequal temperature to neutralize each other and thus produce a current from high to low potential, as when two oppositely charged bodies are connected by a wire.

671. Cause of the Thomson effect. In metals which act like copper, the Thomson effect is said to be positive. The heated end of a bar is electropositive to the cooler end. This is probably due to a diffusion of electrons away from the region where the thermal agitation is greatest, as would be the case with gas in a sealed tube. Most of the gas molecules would collect in the cooler end, like the dots in Fig. 72, which are supposed to represent free electrons.

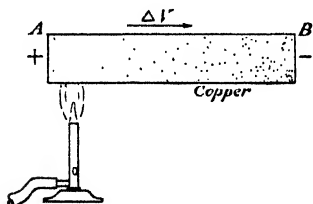


Fig. 72.

In certain metals such as iron, nickel, bismuth, and platinum, the Thomson effect is negative, which means that heat flows more freely against the current than with it. This anomaly can be explained either by supposing that the electrons themselves carry thermal energy, or by supposing that the current is carried more by positive than by negative ions. Both explanations are open to serious objections, so it must be admitted that the negative Thomson effect has not yet been satisfactorily accounted for.

672. Cause of the Peltier effect. The Peltier e.m.f. is closely related to contact difference of potential, and this in turn is closely related to the photoelectric effect and thermionic emission to be discussed farther on. These latter phenomena depend upon the possibility of electrons escaping through the free surface of a conductor either when it is illuminated or heated.

If two metals are placed in contact, and the free electrons from one escape more readily through its boundary than they do from the other, a situation represented in Fig. 73 arises. More electrons have crossed the boundary from metal *A* into metal *B* than have crossed from *B* into *A*. Therefore *B* has acquired a negative charge close to the bounding surface, and *A*, by losing electrons, has acquired a positive charge. These charges attract each other and establish a field adverse to the further escape of electrons at a given temperature. If the junction is heated, more electrons cross the border as a result of thermal agitation, and the difference of potential between *A* and *B* increases. The two charges facilitate the flow of a current from right to left in both bars, and thus constitute a source of e.m.f. directed across the junction from negative to positive, as indicated by the arrow.

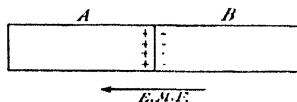


Fig. 73.

We find from the photoelectric and thermionic tables that it takes a higher voltage to stop electrons from escaping from iron than from copper when these metals are illuminated or heated. This amounts to saying that electrons have a greater tendency to leave iron than to leave copper, so that *A* should represent iron in the diagram and *B* copper. Therefore the thermo-e.m.f. must be directed from copper to iron across the junction, as has been determined by experiment.

673. Theory of thermoelectric curve. In order to account for the parabolic form of the exponential curve of thermo-e.m.f. as a function of the temperature, it is necessary to assume that the Peltier effect varies directly as the temperature, and the Thomson effect as the temperature squared. If the cold junction is maintained at 0° C, the equation of the first parabola of Fig. 70 may be written

$$E = At - \frac{1}{2}Bt^2, \quad (1)$$

where *A* and *B* are constants which determine the Peltier and Thomson potentials respectively, and depend upon the metals used. This equation is equivalent to that of a projectile, with *t* representing time and *E* the vertical height *y*, or $y = v_y t - \frac{1}{2}gt^2$, where v_y is the vertical

component of the original velocity and $\frac{1}{2}gt^2$ gives the steadily increasing effect of gravitation.

Just as gravitation tends to neutralize and finally reverse the upward motion of a projectile, so the Thomson effect tends to neutralize the rising thermo-e.m.f. due to heating one junction of a couple. At first its effect is negligible, because B is much smaller than A , but as the Thomson effect varies as the square of the temperature, it increases more rapidly than the Peltier e.m.f., and ultimately reverses it, as we have seen. This may be represented diagrammatically as in Fig. 74.

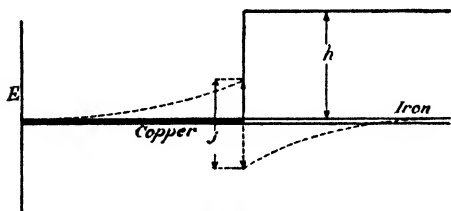


Fig. 74.

The solid line represents the abrupt rise, h , of potential in going across the junction from copper to iron, while the dotted lines show the gradual change of potential along the bars due to the Thomson effect. It is

clear that the latter tends to diminish h by an amount j , because the rise of potential due to the heating of one end of the copper bar is in opposition to the contact difference between it and iron. Similarly, lowering the potential of the iron by heating it, tends to lower its potential with respect to the copper. At the temperature of inversion, $j = h$, and beyond that, when $j > h$, the current flows the other way, or from iron to copper across the heated junction, for then the iron is at a lower potential than the copper.

674. Calculation of thermo-e.m.f. In the general case when the temperature of the cooler junction may have any value, not necessarily 0°C , and when B is not necessarily negative, the equation of the thermo-e.m.f. is

$$E = A(t_2 - t_1) + \frac{B}{2}(t_2^2 - t_1^2), \quad (1)$$

where t_1 is the temperature of the cooler junction and t_2 is the temperature of the heated junction. When $t_1 = 0$, and B is negative, this equation reduces to (1) of Article 673. The values of the constants depend upon the metals used, and are each the algebraic differences of two other constants. Thus $A = a_M - a_N$, and $B = b_M - b_N$, where a and b measure the Peltier and Thomson effects in the metals M and N referred to lead as a basis of comparison. Lead has been taken for this purpose because it has practically no Thomson effect.

The following table contains values of a and b for several metals in terms of microvolts per degree and microvolts per degree squared, respectively.†

Metal	a	b
Iron.....	+ 16.7	$- 3.0 \times 10^{-2}$
Cadmium.....	+ 3.1	$+ 2.9 \times 10^{-2}$
Zinc.....	+ 3.0	$- 1.0 \times 10^{-2}$
Copper (drawn).....	+ 2.8	$+ 1.2 \times 10^{-2}$
Platinum.....	- 3.0	$- 3.2 \times 10^{-2}$
Paladium.....	- 7.4	$- 3.9 \times 10^{-2}$
German Silver.....	- 10.9	$- 3.3 \times 10^{-2}$

If we wish to obtain the equation of a copper-iron couple, we have only to take the algebraic difference of the values of a for the two metals to find A , and of b to find B . Thus the thermo-e.m.f. of a copper-iron couple when $t_1 = 20^\circ$ and $t_2 = 100^\circ$ becomes

$$E = 13.9 (100 - 20) - \frac{0.042}{2} (10,000 - 400).$$

$$\therefore E = 910.4 \text{ microvolts.}$$

675. Thermopiles and pyrometers. The **thermopile** is a series of junctions connected as shown in Fig. 75. If the four junctions a are kept at a constant and relatively low temperature, and the junctions b are warmed, an e.m.f. four times as strong as that due to one junction will be produced. Thermopiles with many couples made by joining fine wires are very sensitive to radiant heat. Their junctions are flattened out into small discs coated with lampblack to increase their power of absorption.

In order to measure temperatures too high for an ordinary thermometer, a single thermojunction **pyrometer** is much used. It is made of two metals joined at a point that is subjected to the temperature to be measured. For very high temperatures one of the wires is of platinum and the other of an alloy of platinum and rhodium. A baked clay inner tube like a pipe stem separates the two wires just above the junction, as shown in Fig. 76, and a larger porcelain tube

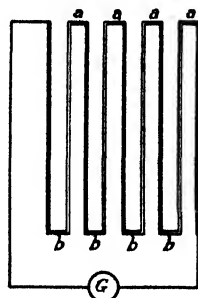


Fig. 75.

† Taken from the *Handbook of Chemistry and Physics*, Chemical Rubber Publishing Co.

protects the junction and the wires from too intimate contact with the molten metal of the furnace.

If the wires from the pyrometer to the milliammeter and its own winding are of copper, there are two junctions at room temperature instead of one. But there is a principle of thermoelectricity according to which the difference between the two effects at K and K' is the same as the single e.m.f. that would exist if the platinum wire were joined directly to the alloy at the same temperature. So the meter is affected only by the difference between the temperature of the junction J and the temperatures of K and K' , which are supposed to be the same.

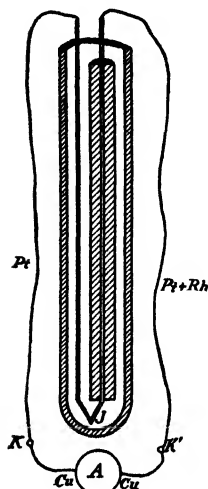


Fig. 76.

676. The thermocouple meter. Ordinary galvanometers do not detect alternating currents because of the periodic reversal of the torque. But if the current, whatever its frequency, is sent through a fine wire of high resistance, the wire is heated. The resulting rise of temperature may heat a thermojunction whose thermo-e.m.f. causes a direct current to flow through an ordinary galvanometer. An instrument of this kind is shown

in its essential parts in Fig. 77, where J is a thermojunction in contact with a fine wire stretched from a to b . Through this wire is sent the alternating or oscillatory current to be measured, and so the wire is heated. This heats the junction J , and the resulting thermo-e.m.f. sends a current through the galvanometer G . The deflection of the galvanometer depends upon the heat H , which is proportional to I^2R , according to Joule's law. Thus the deflections depend upon the current squared. They would vary *directly* as I^2 if the thermo-e.m.f. were a linear function of the temperature. But the thermo-e.m.f. increases at a decreasing rate with rising temperatures, as we have seen. This tends to offset the increasing effect due to the current squared. However, as the rise of temperature never approaches the value at which the thermo-e.m.f. begins to decrease, the thermogalvanometer, on the whole, becomes increasingly sensitive with increasing current, and the scale divisions are therefore more widely spaced for the larger current values.

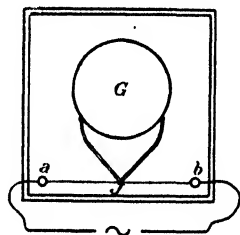


Fig. 77

SUPPLEMENTARY READING

- A. W. Hirst, *Electricity and Magnetism* (Chap. 7), Prentice-Hall, 1937.
C. A. Culver, *Electricity and Magnetism* (Chap. 13), Macmillan, 1930.
Page and Adams, *Principles of Electricity* (pp. 218-233), Van Nostrand, 1931.

PROBLEMS

1. Using the values of a and b in Article 674, calculate the thermo-e.m.f. of a copper-iron couple when one junction is at 0°C and the other at 200°C .
Ans. 1940 microvolts.
2. Calculate the temperature of inversion and neutral temperature of a copper-iron couple when the cooler junction is at 0°C . (Use equation (1) of Article 674.) *Ans.* 662°C , 331°C .
3. Calculate the thermo-e.m.f. of a copper-iron couple when one junction is at 50°C , and the other at 200°C . *Ans.* 1297.5 microvolts.
4. Calculate the thermo-e.m.f. of a palladium-zinc couple when one junction is at 0°C , and the other at 100°C . *Ans.* 1185 microvolts.
5. Calculate the thermo-e.m.f. in Problem 4, if the cooler junction is kept at 20°C . *Ans.* 971.21 microvolts.
6. Calculate the neutral temperature of a zinc-iron couple and of a platinum-copper couple. *Ans.* 685°C ; -132°C .

CHAPTER 50

Electrical Measurements

677. Resistances. An indispensable part of the apparatus needed in making most electrical measurements are standard resistances and variable resistances, or rheostats. The former are made by winding bobbins with coils of wire made of an alloy whose resistance is but little affected by temperature. A usual form of resistance box is the

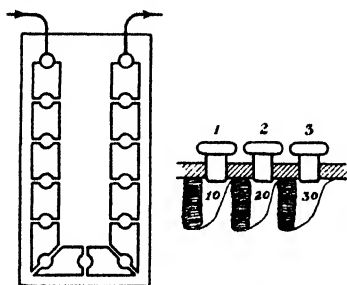


Fig. 78.

box encounters a resistance of 10 ohms. If the first and second plugs are removed, 30 ohms are introduced into the circuit. A removal of the first and third plugs introduces 40 ohms, and so on. Another and increasingly popular type is the dial rheostat built on the decade principle. Such a resistance box may have four dials, one in steps of single ohms, one in tens, one in hundreds and one in thousands. There are ten steps of each kind, so that by rotating the moving arm over the dial, resistances from 1 to 9999 ohms are introduced into the circuit.

There are also many cases in which resistances are used to control the current or vary the potential difference between terminals of a circuit, when the value of the resistance need not be known.

These variable rheostats are of several types. A very convenient one is the drum rheostat, shown in Fig. 79. The coil of high-resistance

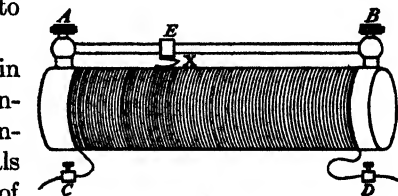


Fig. 79.

wire is wound on an enameled drum and brought out to two binding posts, CD . A sliding contact runs on a conducting rod whose ends are fitted into binding posts, AB . If the circuit is made between A and C , the resistance is due to that portion of the coil between C and X .

Another convenient form of control rheostat is obtained by having coils of different resistances connected as shown in Fig. 80. On 100 volts, with the first switch closed, half an ampere flows. The second switch gives one ampere, the first plus the second gives 1.5 ampere, the third gives 2 amperes, the first plus the third gives 2.5 amperes, and so on in stages of half an ampere up to 15.5 amperes, when all the switches are closed. If all are open, the resistance is infinite and there is no current.

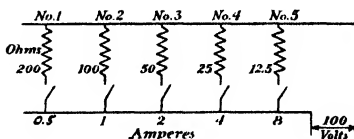


Fig. 80.

678. Galvanometers. Oersted's classic experiment with the compass under a wire carrying a current represents a galvanometer, because the deviation of the compass depends upon the strength of the current. Such a galvanometer can therefore be used as an indicator both of current strength and direction. But the same result could have been obtained if the magnetized compass needle had been held stationary, and the wire had been delicately pivoted and so capable of motion. The force that acts between them would have moved the wire instead of the needle, as we should expect from the discussion in Article 621. Thus we may have two kinds of galvanometer, one where the wire, or coil, is fixed and the magnet moves, and one in which the magnet is fixed and the coil moves.

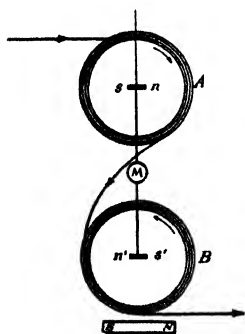


Fig. 81.

679. The fixed-coil galvanometer. The most sensitive instruments are of this type, and are best represented by the Thomson galvanometer, shown diagrammatically in Fig. 81. A group of very short needles sn is at the center of a coil A of many turns of wire. A similar group of needles $n's'$ of reverse polarity is at the center of a similar coil B . Both sets of needles are mounted on a light but rigid rod which is suspended by a fine silk fiber at its upper end. Such a system is called *astatic*, because the directive torque of the earth's field is neutralized by the two sets of needles. The bar magnet SN below the coil B supplies a directive

torque by its action on $n's'$, so that the moving system has a definite zero position when no current is flowing.

The winding sense of the two coils is opposite, so that the torques on the two sets of needles are in the same sense when the current is flowing. The resulting deviation of the suspended system is measured by a beam of light reflected from the mirror M upon a scale, or by observing the scale as seen reflected in the mirror with the aid of a telescope. This Thomson galvanometer may be so sensitive as to deflect a beam of light through a measurable angle when the current is less than 10^{-10} ampere. But it has the serious disadvantage of being somewhat affected by stray magnetic fields.

680. The moving-coil galvanometer. This instrument is the reverse of the galvanometer described in the last article. Instead of a fixed coil acting on a light movable magnet, a fixed and heavy magnet acts upon a light movable coil. It is often called a D'Arsonval galvanometer, a name derived from that of its inventor.

The magnet in typical D'Arsonval galvanometers is of the horse-shoe type, and is usually fitted with curved pole pieces so as to concentrate the field. This field is normal to the central

turns of the coil as it rotates about a vertical axis within the angle α , as shown in Fig. 82 (b). The coil consists of many turns of fine wire, and the current is led in by a narrow strip of phosphor bronze or gold, which

acts as the suspension shown in (a). The circuit is completed through a helical spring of the same material as the suspension, and so flexible as to produce a negligible torque.

When a current flows

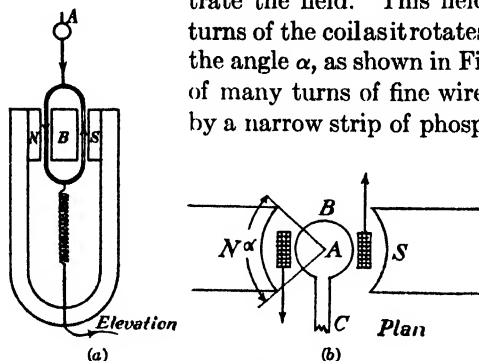


Fig. 82.

through the coil, if its direction is downward in the left half and upward in the right half, two forces shown by the arrows in (b) act upon it, producing a torque around the axis A . An iron cylinder B , supported by a horizontal rod C , helps to strengthen the field and therefore the torque, and also to make the lines of force more nearly radial where the coil cuts them. If they were exactly radial, the torque would be directly proportional to the current. The restoring torque supplied by the phosphor bronze suspension is proportional to the angular twist of the coil; therefore, when there is equilibrium between the two torques,

the angular deviation is directly proportional to the current and to the field strength. This is the fundamental principle of the instrument. However, as it is impossible to calculate the value of the field strength, the constant of the galvanometer must be determined by experiment. This "constant," unfortunately, changes slowly over a period of years as the magnet grows weaker with age.

Since in a well-made moving-coil galvanometer, the angular deviation θ varies as the current, it follows that $I = k\theta$, where k is the constant of the instrument. We can determine k by sending a very small current through the coil, and observing the resulting angular deviation. In practice, θ is observed indirectly by reflecting a pencil of light from the mirror to a scale over which the "spot" travels as the coil turns. If the scale is at the standard distance of one meter, the current in amperes necessary to produce one millimeter of deflection is known as the **figure of merit**. This constant may be as small as 2×10^{-10} ampere, or one five thousandth of a microampere per millimeter. This high sensitivity demands a coil with many turns, a very concentrated field, and a very delicate suspension.

681. The ammeter. In measuring currents such as occur in commercial practice, the galvanometer is far too sensitive. But it may be used as an indicator, provided only a small fraction of the main current is passed through it. To accomplish this, use is made of a low-resistance *shunt* across the galvanometer terminals. Further, as very great sensitivity is no longer necessary, the galvanometer may be made portable. This is accomplished by having the armature pivoted, and by using fine spiral springs to carry the current and supply the restoring torque. Instead of a mirror and scale, a light pointer fastened to the coil, or *armature*, turns with it and passes over a scale divided to read amperes or milliamperes directly, as shown in Fig. 83(a). The electrical connections are shown diagrammatically in Fig. 83(b), where the shunt S is a bar or strip of metal of very low resistance, while that of the winding of the galvanometer is relatively high. The current divides at a inversely as the resistances (see Article 633), so that if the galvanometer has a resistance of 99.9 ohms, and the shunt 0.1 ohm, 0.001 of the total current flows through the

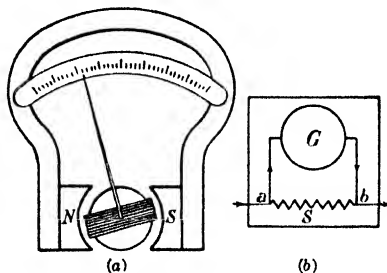


Fig. 83.

resistance, while that of the winding of the galvanometer is relatively high. The current divides at a inversely as the resistances (see Article 633), so that if the galvanometer has a resistance of 99.9 ohms, and the shunt 0.1 ohm, 0.001 of the total current flows through the

galvanometer, and 0.999 through the shunt. As this ratio is independent of the current, the deflections of the galvanometer are very nearly proportional to the main current. The scale may then be calibrated in nearly equal divisions reading amperes directly.

682. The hot-wire ammeter. In measuring high-frequency currents such as are used in radio transmission, we may use either a meter

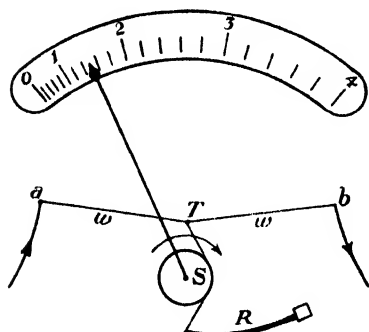


Fig. 84.

with a thermojunction, such as was described in Article 676, or a hot-wire meter. The essential parts of this instrument are shown in Fig. 84. A fine wire w is fastened to the supports a and b , by which it becomes part of the circuit. A thread fastened to the wire at T passes around a spindle S and to a spring R , which keeps it tight and takes up the slack in the wire. The pointer P turns with the spindle and indicates the current. When the current passes, the wire gets hot and expands, the thread takes up the slack, and the spindle rotates clockwise. As the heat developed varies with I^2 , the deflections vary approximately with I^2 also. This results in making the higher scale divisions farther apart than the lower ones, so that the instrument grows increasingly sensitive as the current increases.

683. The voltmeter. A pivoted moving-coil galvanometer may be used to indicate differences of potential as well as current strength. In this case it must be protected by a high resistance from the destructive current which would flow through its fine winding if commercial voltages were impressed upon it. This resistance also serves to bring its readings within the range desired. Both galvanometer and auxiliary resistance are usually enclosed within the same box, as indicated in Fig. 85.

The calculation of the resistance for a given range of voltages is as follows: Suppose a pivoted galvanometer whose needle sweeps over the entire scale with a current of 100 milliamperes, and it is desired to have the instrument read to 150 volts. Then the combined resistance of galvanometer and protecting coil is given by $R + r = E/I = 150/0.1 = 1500$ ohms. If the galvanometer resistance r is 100 ohms,

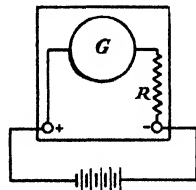


Fig. 85.

R must be 1400 ohms. As the deflections are nearly proportional to I , and as $R + r$ is constant, they are also proportional to E . The scale may then be laid off in nearly equal divisions reading volts directly.

There are two uses for the voltmeter: in the first case, as indicated in Fig. 85, it is connected directly across the terminals of a battery or generator. It then reads the terminal voltage, which is always less than the e.m.f. because of the internal drop. However, as the voltmeter draws very little current, this loss is small in most cases for which it is proper to use a voltmeter at all.

The other use is in determining the fall of potential across a resistance r carrying a current I , as indicated in Fig. 86. The galvanometer current i equals Ir/R , since $i/I = r/R$. Therefore the deflection of the needle is a measure of the fall of potential Ir across the resistance r . This gives a nearly correct value for the drop before the voltmeter was connected, provided the resistances of the battery and leads, as well as r , are small compared to the resistance of the voltmeter. If they are not, the increase in the main current due to connecting the

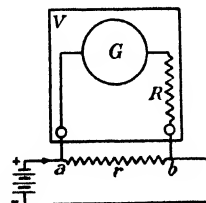
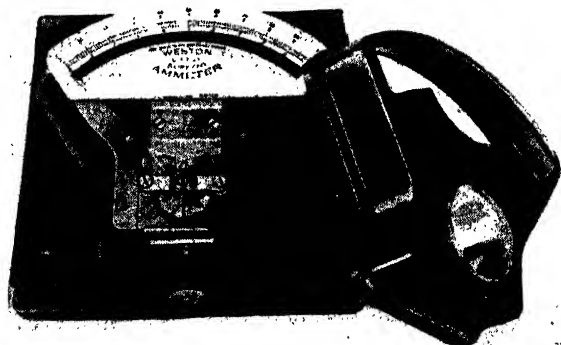


Fig. 86.



Courtesy Weston Electrical Instrument Co.

Plate 16.

Photograph illustrating essential parts of portable
D.C. ammeter.

voltmeter in parallel with r results in a decrease in the potential drop that was to be measured. A correction of the reading is then necessary in order to obtain the value desired.

684. Wattmeters. If the power delivered to a circuit is to be measured, we may connect a voltmeter across the terminals of the

load, as indicated in Fig. 87, and an ammeter in series with the circuit. This enables us to calculate the power in watts by the familiar relation $P = EI$. But it is more convenient to have a single instrument

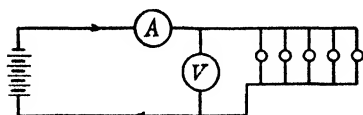


Fig. 87.

which reads watts directly. This is accomplished by using a galvanometer, whose field is not supplied by a permanent magnet but by a coil of a few turns of heavy wire capable of carrying the entire current

I , or a determined fraction thereof. Within this field is an armature of fine wire protected by a high resistance in series with it, as in a voltmeter. The flux, indicated by the arrows in Fig. 88, sets up a torque in the armature exactly as is the case when a permanent magnet is used. But here the field is not constant. It varies directly as the current I , just as the field H varies at the center of the coil of a fixed-coil galvanometer. The torque depends upon both the field strength and the current i flowing in the armature, so we may write $L \propto iH$. But $i = E/R$, as in the case of the voltmeter, where R is the total resistance of the armature circuit. Then since $H \propto I$ we may substitute these values in the first variation, and obtain

$$L \propto \frac{EI}{R}.$$

Thus since R is a constant, the torque is a direct measure of EI ,

or the power delivered to the load, and the scale may be laid off in nearly equal divisions to read watts directly.

As one terminal of the armature circuit is permanently connected to one terminal of the field circuit at c , only three external terminals are needed, namely, a , b , and c , of which a and c are in series with one side of the line, while b is connected to the other.

Such instruments read the power of alternating as well as of direct currents, because if the current changes sign, both field and armature currents reverse simultaneously and the torque is still in the same direction.

685. The electrodynamicometer. In general, all meters like the foregoing, in which an interaction between two circuits results in a torque that may be measured, are called *electrodynamometers*.

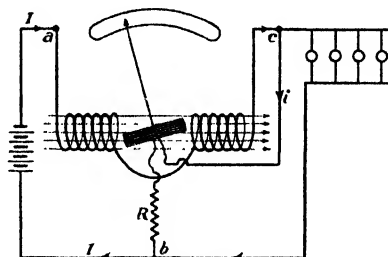
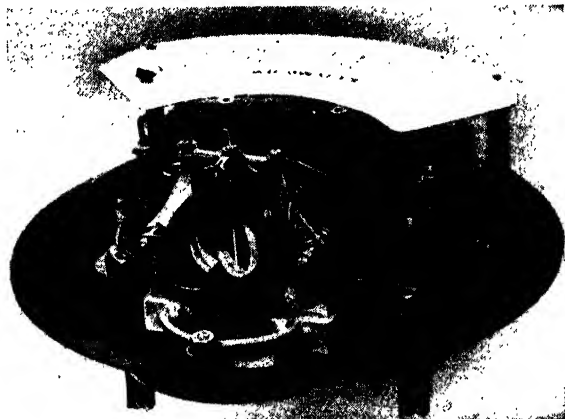


Fig. 88.

Both voltmeters and ammeters may be constructed on this principle, and thus be made available for alternating currents as well as direct. In the case of the latter, the same current flows through both the fixed



Courtesy Weston Electrical Instrument Co.

Plate 17.

Photograph of a wattmeter, showing two fixed coils and the moving coil (white) mounted on an axis to which the needle is attached. The plates in the rear include the resistors of the potential circuit, indicated by R in Fig. 88.

coil, which may be regarded as establishing the field, and through the movable coil, which represents the armature. This is shown diagrammatically in Fig. 89, where A is the armature supported by pivots (not shown), and F is the field coil. The torque L , as usual, is proportional to IH , but since H varies as I , we may write $L \propto I^2$. Then the current may be found from the torque that must be set up in a helical spring acting upon the armature in order to keep it from turning; or a pointer attached to the armature may move over a scale whose divisions are spaced so as to read amperes directly.

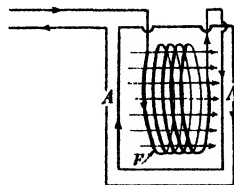


Fig. 89.

686. Electrostatic voltmeters. These instruments are especially useful in reading potentials too high for a dynamometer type of meter. There are many kinds, but all depend upon the electrostatic forces between charged bodies. In Fig. 90, the plates PP are connected to one terminal of the source of e.m.f., and the needle N , pivoted at A , to the other. If P is either positive or negative, N will always have a

charge of opposite sign, and is attracted toward P , causing the pointer to sweep over the scale. In this form of the instrument, the restoring torque is supplied by gravity acting upon the needle, whose center of mass is above the horizontal axis. The small weights w

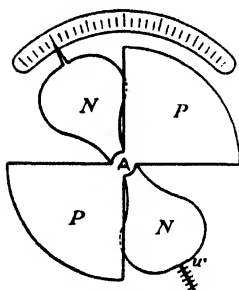


Fig. 90.

partially counteract this torque, and may be adjusted for purposes of calibration. As the charges on plates and needle are directly proportional to the charging e.m.f., the deflection, when equilibrium is attained, is a direct measure of the applied voltage.

687. The potentiometer. When it is desirable to measure an electromotive force with extreme precision, a device known as *Poggendorf's potentiometer* is used instead of a voltmeter. This consists in comparing an unknown with the known e.m.f. of a standard cell. The fundamental principle of the potentiometer is shown in Fig. 91. There are two circuits. The main circuit contains a storage cell B , whose e.m.f., E , must be greater than those used in the branch circuit. The other contains either the standard or unknown cell S or X , whose electromotive forces are E_s and E_x . In the main circuit Bdc , the fall of potential across dc is adjusted to equal and oppose the e.m.f. of the standard cell or unknown cell in the shunt circuit Sdc (or Xdc). Then no current can flow through the

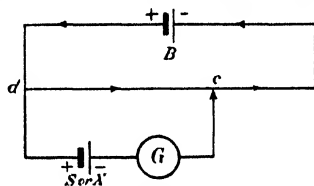


Fig. 91.

shunt circuit and the galvanometer G is not deflected.

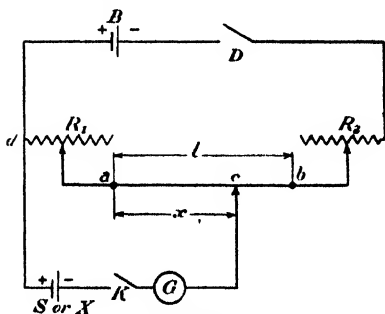


Fig. 92.

Figure 92 illustrates the connections in more detail. The variable resistances R_1 and R_2 control the current through a wire ab whose resistance and length are r and l . Any change in R_1 must be compensated by a corresponding change in R_2 , so that the total resistance, and therefore the

main current, may remain constant. A sliding contact c adjusts the potential drop so that the potentiometer may be "balanced."

This balance is shown by the galvanometer, which then does not deflect when the key K is closed.

Let us suppose that with the standard cell in the branch circuit, the galvanometer does not deflect when the contact is x centimeters from c . Then the resistance of the wire between c and a is rx/l , and the total shunted resistance is $R_1 + rx/l$. When the unknown cell is connected, there are, in general, different values of R_1 and x that we may indicate by R'_1 and x' , so the shunted resistance is now $R'_1 + rx'/l$. As the main current has been maintained constant, the ratio of the two voltages is equal to the ratio of these shunted resistances, or

$$E_x = E_s \frac{R'_1 + rx'/l}{R_1 + rx/l}. \quad (1)$$

If real precision is not needed, we may dispense with R_1 and R_2 , making the balance wholly with the wire. Then equation (1) reduces to the very simple formula

$$E_x = E_s \frac{x'}{x}, \quad (2)$$

where it is not necessary to know either the resistance or length of the wire, but only the length of the two segments x and x' .

688. Wheatstone's bridge. The usual measurements of resistance depend upon a comparison of the resistance to be measured with a standard resistance. The best-known method is due to Wheatstone,† who made use of the network shown in Fig. 93. A battery of voltage E sends a current to the junction at a , where it divides, part flowing through the unknown resistance X and then through the known resistance R , and part through the resistances A and B . If the four resistances are so adjusted that the fall of potential through A is equal to that through X , then a galvanometer connected between b and d will not deflect when the key K is closed. This is because b and d are at the same potential level. The same would be true of the flow of water if the figure $abcd$ were an island with a river flowing

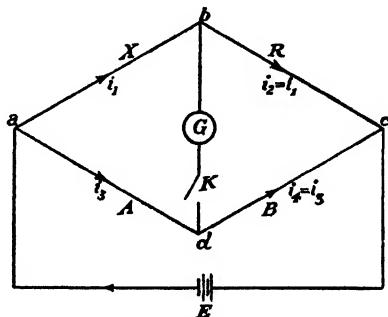


Fig. 93.

equal to that through X , then a galvanometer connected between b and d will not deflect when the key K is closed. This is because b and d are at the same potential level. The same would be true of the flow of water if the figure $abcd$ were an island with a river flowing

† Charles Wheatstone (1802-1875), an English physicist and inventor.

around it. A canal cut across between two points where the level of the branches of the stream was exactly the same would have no current.

If we assume the bridge to be balanced, with no current flowing through G , $i_1 = i_2$, $i_3 = i_4$, and the potential differences are equal in pairs, or

$$i_1 X = i_2 A \quad (1)$$

and

$$i_3 R = i_4 B. \quad (2)$$

Dividing (1) by (2) gives

$$\frac{X}{R} = \frac{A}{B},$$

or

$$X = R \frac{A}{B}, \quad (3)$$

from which X can be calculated if R and the ratio of A to B are known.

689. The slide-wire bridge. There are two types of the Wheatstone bridge in common use. Both are based on the same principle, but it is differently applied in each. In the slide-wire pattern, the resistances A and B are replaced by a wire of german silver or other high-resistance alloy, and a sliding contact which is opened or closed

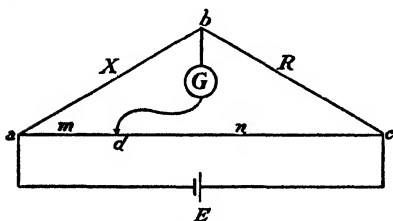


Fig. 94.

at will connects the galvanometer to the movable point d , as in Fig. 94. If the wire is of uniform cross section, the resistances of the segments m and n (measured in centimeters) are proportional to their lengths. Therefore $A/B = m/n$, where A and B are the resistances

of the segments. The bridge equation now becomes $X = Rm/n$. In this case it is evident that the resistances A and B need not be known, as the ratio $m : n$ is that of two lengths read in centimeters or any other convenient unit.

690. The post-office bridge. This (the second type) is the form most commonly used in both laboratory and commercial testing. It is so named because it was first developed for use by the British Post Office in connection with the telegraph. It differs from the slide-wire pattern of Wheatstone's bridge in having a *fixed* ratio $A : B$, but with R variable. Such bridges are usually made in a portable box with all the parts, including battery and galvanometer, assembled in a compact form. The resistance coils representing either A or B may each be chosen from four units wound to 1, 10, 100, and 1000 ohms,

so that ratios of 1 : 1, 1 : 10, 1 : 100 and 1 : 1000, as well as their reciprocals, may be obtained. Thus if a 10-ohm coil is chosen for A , and 1000 ohms for B , the ratio $A : B$ equals 0.01, and X is one-hundredth part of R .

691. Condensers. Measuring the capacitances of condensers is an important part of laboratory procedure. In any condenser, parallel plates separated by a dielectric have a capacitance given by $C = Q/\Delta V$, where Q is the charge on either plate, and ΔV the difference of potential between them. Such a condenser may be adapted to the relatively low values of ΔV obtained from batteries and generators by multiplying the number of the plates and consequently increasing the total area. This greatly increases the capacitance, so that even if

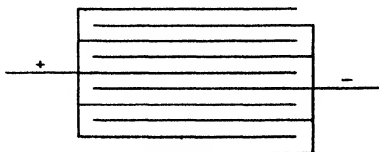


Fig. 95.

ΔV is small, Q may reach any desired value. The condenser, shown diagrammatically in Fig. 95, is made of a great number of interleaving sheets of tin foil separated by thin layers of paraffin, or still better, of mica. Each side has a total area of n times the area A of each sheet, where n is the number of sheets on each side. The total capacitance is therefore $2n - 1$ times as great as that of a single pair, as seen in the diagram, where there are ten sheets, but only nine condensers.

692. Dimensions and unit of capacitance. As we have seen, the dimensions of capacitance in the electrostatic system are those of length. In the electromagnetic system we may obtain the dimensions of C from the defining equation $[C] = [Q/V]$; hence, substituting the dimensions of these quantities, we obtain

$$[C] = [M^{\frac{1}{2}}L] \div [M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-2}] = [L^{-1}T^2].$$

The ratio of the dimensions of C in the two systems is therefore

$$\left[\frac{C_e}{C_m} \right] = \left[\frac{L}{L^{-1}T^2} \right] = [L^2T^{-2}].$$

The quantity L^2/T^2 is the square of a velocity, and accurate measurement shows that it is the velocity of light, c . This illustrates the important fact discovered by Maxwell that the ratio of the electrostatic to the corresponding electromagnetic units is equal to c or one of its powers as shown for currents in Article 617.

The absolute unit of capacitance in the c.g.s. system is that capacitance which is charged to one absolute unit of quantity by the abso-

lute unit of potential difference. This applies to both the e.s.u. and e.m.u. systems. But if we use the coulomb (one-tenth of the e.m.u.) and the volt (10^8 times the e.m.u.), we obtain the corresponding capacitance unit which is called the **farad**, after Michael Faraday. Therefore the farad is 10^{-9} of the absolute unit in the electromagnetic system. Even this unit is much too large for common practice, and the **microfarad**, one millionth of the farad, and 10^{-15} of the absolute unit, is the usual measure of capacitance.

693. Calculation of the capacitance of a condenser. If the value of the capacitance of a parallel-plate condenser, as obtained in equation (7), Article 602, is multiplied by n , the number of pairs of leaves, we obtain

$$C = \frac{KnA}{4\pi t},$$

where A is the area of each leaf in square centimeters, t is the thickness of the dielectric in centimeters, and K is the dielectric constant. To reduce this to absolute electromagnetic units, we must divide by the square of the velocity of light, or 9×10^{20} cm/sec. But the microfarad is 10^{-15} of the absolute unit; therefore C measured in this smaller unit is numerically 10^{+15} times larger, giving

$$C = \frac{KnA \times 10^{15}}{4\pi t \times 9 \times 10^{20}} = \frac{885KnA}{10^{10}t} \text{ microfarads,} \quad (1)$$

which is a convenient formula for calculating the capacitance of commercial condensers.

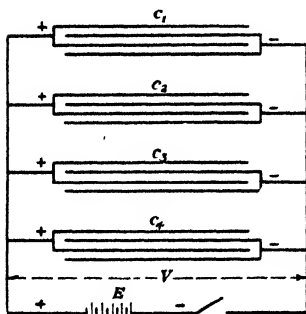


Fig. 96.

694. Combinations of condensers.

In order to obtain a large capacitance suitable for the relatively low potentials of batteries and generators, several condensers may be connected in parallel, as in Fig. 96. In this case it is almost self-evident that the total capacitance, C , is equal to the sum of their individual capacitances, $C_1 + C_2 + \dots + C_n$. This, however, may be proved as follows: The total charge is equal to the sum of the charges on each condenser,

or $Q = Q_1 + Q_2 + Q_3 + \dots + Q_n$. Therefore, since $Q = VC$,† and as

†In this discussion, V is written in place of the more correct ΔV , to avoid unnecessary complexity.

each condenser is charged to the same potential V , we may substitute for the values of Q , and obtain

$$VC = VC_1 + VC_2 + VC_3 + \dots VC_n.$$

$$\therefore C = C_1 + C_2 + C_3 + \dots C_n.$$

The condensers used in connection with low-voltage circuits are unsuitable for use with high potentials, and there is always a maximum allowable potential which should not be exceeded. This varies according to the nature of the dielectric and its thickness, being as low as 40 volts for some paraffin condensers, and 150 volts for many made of mica. In order to use such condensers with higher voltages, it is necessary to connect them in series, as in Fig. 97, so that each will

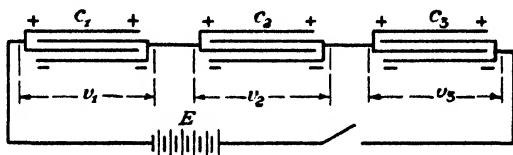


Fig. 97.

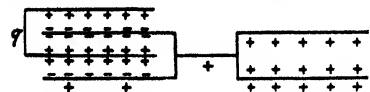
have only a part of the total e.m.f. impressed upon it. In this case it is obvious that $V = v_1 + v_2 + v_3 + \dots v_n$, where $v_1, v_2, \dots v_n$ are the potentials across the terminals of the individual condensers. But the charge Q is about the same for each, since if $+Q$ units are on one set of plates of any condenser, $-Q$ units must be on the other set, according to the principles established by Faraday's ice-pail experiment. Consequently there is a charge $+Q$ on every alternate set of plates, and a charge $-Q$ on those in between.† In other words, there is a uniform *displacement* of Q units of electricity throughout the circuit. Therefore, we may substitute $V = Q/C$ for the total potential, and $v_1 = Q/C_1$, $v_2 = Q/C_2$, and so on, for the various partial potentials, and obtain

$$\frac{Q}{C} = \frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3} + \dots \frac{Q}{C_m}.$$

Then, dividing by Q , we have

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots \frac{1}{C_m},$$

† This is not strictly true, because not quite all of the repelled charge appears on the next condenser. A very small portion of it is on the connecting wire, and on the outer surface of one of the two outer sheets of the condenser from which it is repelled.



This is indicated in the accompanying drawing, where the number of $+$ and $-$ signs indicates the relative magnitudes of the charges.

from which C , the total capacitance of condensers in series, may be approximately calculated in the same manner as the total resistance of a number of resistances in parallel.

SUPPLEMENTARY READING

- A. Zeleny, *Elements of Electricity* (Chapters 11, 21), McGraw-Hill, 1930.
C. A. Culver, *Electricity and Magnetism* (Chap. 7), Macmillan, 1930.
A. W. Smith, *Electrical Measurements in Theory and Applications*, McGraw-Hill, 1934.

PROBLEMS

1. A current of 3.2×10^{-7} ampere causes the "spot" of a D'Arsonval galvanometer to deflect 16.8 cm on a scale 75 cm from the mirror. What is its figure of merit? *Ans.* 1.43×10^{-9} ampere per millimeter.

2. A portable galvanometer, whose needle deflects 5 scale divisions per milliampere, is to be used in an ammeter. Its resistance is 238 ohms. What should be the resistance of the shunt in order that the needle may deflect 10 divisions per ampere? *Ans.* 0.477 ohm.

3. It is desired to use the galvanometer of Problem 2 in a voltmeter to read half a volt per division. What resistance must be connected with it in series? *Ans.* 2262 ohms.

4. A voltmeter whose internal resistance is 90 ohms is used to read the drop of potential across 45 ohms in a circuit whose total resistance is 75 ohms, when an e.m.f. of 150 volts is impressed upon it. What is the drop to be measured, and what is the observed value? *Ans.* 90 volts; 75 volts.

5. A potentiometer uses a standard cell whose e.m.f. is 1.0185 at 15°C . The slide wire is 1 m long, and has a resistance of 1.50 ohm at 15°C . When used to measure the e.m.f. of an unknown cell at this temperature, the resistances R_1 and R_1' are 8 and 12 ohms respectively. The distances x and x' are 36.2 and 64.3 cm, respectively. Calculate E_x . *Ans.* 1.546 volt.

6. In a Wheatstone slide-wire bridge the balance is obtained at a point 32 cm from one end of a wire 120 cm long. The unknown resistance is connected to that end, and the standard resistance is 4.5 ohms. What is the value of X ? *Ans.* 1.64 ohm.

7. Calculate the capacitance of a condenser made of 2 sets each of 6 interleaving semicircular metallic sectors whose radii are 6 cm, with an air space of 2 mm between them. *Ans.* 0.275 millimicrofarad.

8. Calculate the capacitance of a condenser made of 2 sets each of 150 tin foil sheets measuring 20×30 cm and separated by layers of mica 0.1 mm thick, and whose dielectric constant is 6. *Ans.* 9.53 microfarads.

9. What is the capacitance of 6 condensers each of 0.2 microfarads capacitance when connected in parallel? In series? *Ans.* 1.2 m.f.; $0.033 +$ m.f.

10. Three condensers, each of 0.2 microfarad capacitance, are raised to a potential of 110 volts. What charges do they receive when connected in series and in parallel? *Ans.* 7.33×10^{-6} coulomb; 66×10^{-6} coulomb.

CHAPTER 51

Electromagnetism

695. The solenoid. This is a helically wound coil of wire through which a current may be passed, thus producing a magnetic field of force which interlinks with the current. A solenoid is shown in Fig. 98, where the dotted lines represent the field. The general effect is the same as in the case of a single turn, but the magnitude of the field and its distribution are quite different, so that a new equation must be derived to meet the altered conditions.

In order to solve this problem, we apply Laplace's formula, and with the aid of the calculus find that the field at the center of a solenoid is given very nearly by

$$H = \frac{4\pi nIl}{\sqrt{l^2 + 4r^2}}, \quad (1)$$

where l is the length of the solenoid, r is its radius, n is the number of turns per unit length, and I is given in e.m.u.

There are two important special cases to which equation (1) may be applied, and which reduce it to a simpler form. If the solenoid is very long compared to its radius—that is, $l \gg r$ —then $4r^2$ may be neglected compared to l^2 , and H becomes $4\pi nI$. If I is measured in amperes,

$$H = 0.4\pi nI,$$

or

$$H = 0.4\pi NI/l, \quad (2)$$

where N is the total number of turns. If l is ten times the diameter of the solenoid, this equation is only about 0.5 per cent in error at the center. At the ends of a long solenoid, H is one half of its value at the center, as there is a continuous sidewise leakage of lines that never reach the ends.

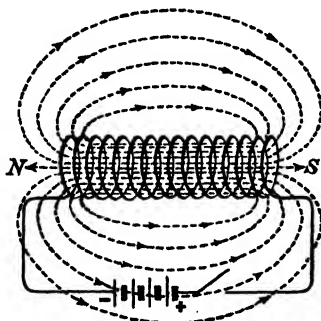


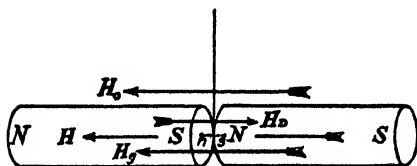
Fig. 98.

The other case is when $r \gg l$. Then H becomes $2\pi nIl/r$, or

$$H = \frac{2\pi NI}{10r}, \quad (3)$$

when I is in amperes. This is the field at the center of a narrow coil such as is used with moving-needle galvanometers. When $N = 1$, we recover the familiar equation for H at the center of a single turn.

696. Magnetic flux in a field of force. If an iron core is introduced into a solenoid through which a current is flowing, the magnetic field induces poles at the ends of the core corresponding to those of the solenoid itself. This effect is accompanied by an increase of the field within the iron. In order to study the changed conditions, let us suppose that the core is cut across at its center, and the two halves separated just enough to permit a miniature magnetized needle to



be swung in the "crevasse" between them as in Fig. 99.

If the crevasse is very narrow, it does not appreciably alter the field, and the needle is acted upon by the resultant magnetic force at the center of the solenoid as if the iron

were not cut in two. This resultant is the net effect of three fields of force. These are: the original field H_0 before the iron was introduced; the field H_g caused by the poles on either side of the crevasse which appear when the core is divided, as explained in Article 562; and the demagnetizing field H_d caused by the free poles at the ends of the core. Calling the total effect B , we have

$$B = H_0 - H_d + H_g. \quad (1)$$

But H_g equals $4\pi\mathcal{Q}$, where \mathcal{Q} is the intensity of magnetization of the core, as was proved in Article 563, equation (3); and we may set $H_0 - H_d = H$ where H is the net field *within the iron*. Then

$$B = H + 4\pi\mathcal{Q}. \quad (2)$$

With a long slender solenoid and core, the demagnetizing field H_d becomes so small that the original and internal fields are sensibly equal, and $H = H_0$. But with short thick bars, as stated in Article 564, H is very much weaker than H_0 and can be calculated only approximately. These bars make poor permanent magnets, because when the

magnetizing field H_0 is withdrawn, $H = -H_D$, and $B = 4\pi\mathcal{Q} - H_D$, where H_D is much greater than in long slender bars.

No matter how short the solenoid, the resultant B is always in the same direction as H , and therefore the lines representing it are directed from S to N inside the iron, and from N to S outside, forming continuous loops, like those shown in Fig. 98, instead of starting at a north pole and ending at a south pole. The lines representing B were called by Faraday **lines of induction**, and, taken altogether, **the magnetic flux**, represented by ϕ . Thus B measures the flux density, or lines of induction per cm^2 . Its unit is the **gauss**,[†] or one line per cm^2 . The total flux ϕ is measured in **maxwells**;[‡] one line of induction is one maxwell. This flux is purely imaginary, for nothing flows, and a field of induction has not the same reality as a field of force. However, the idea of a continuous flux is extremely useful in certain practical problems concerned with magnets and with electromagnetic induction. In fact, Faraday devised the notion of flux in order to formulate the phenomena of induced currents, considered in the next chapter.

697. Permeability and susceptibility. In order to calculate his imaginary flux in a simple manner, Faraday regarded it as a sort of consequence of the field of force in which the iron was placed, somewhat as strain may be regarded as the result of stress. This means that B varies as H , as is evident from equation (2) of the last article. Thus we may write

$$B = \mu H, \quad (1)$$

where μ is a constant of proportionality. This constant is known as the **permeability** of the iron, and it is much used by engineers in designing electrical machinery. Its value in a vacuum is arbitrarily taken as unity, so that $B = H$ in a vacuum, and nearly so in air.

To obtain a relation between permeability and magnetic intensity, we may divide equation (2) of the last article by H , giving

$$\frac{B}{H} = \mu = 1 + \frac{4\pi\mathcal{Q}}{H}. \quad (2)$$

The ratio \mathcal{Q}/H is known as the **susceptibility** of the iron, and is indicated by the Greek letter κ (kappa). So (2) becomes

$$\mu = 1 + 4\pi\kappa. \quad (3)$$

[†] Named for Karl F. Gauss (1777–1855), a celebrated German mathematician.

[‡] Named for James Clerk Maxwell (1831–1879), a celebrated Scottish mathematical physicist.

As permeability is unity in a vacuum, the susceptibility there is zero. With diamagnetic bodies whose permeability is less than unity, κ is slightly negative. With ferro- and paramagnetic bodies it is more or less positive.

The susceptibilities of paramagnetic bodies at 18°C range from the very high figure of 22×10^{-6} for the rare element erbium, 15×10^{-6} for cerium, and 9.9×10^{-6} for manganese, down to such low values as 0.025×10^{-6} for tin. Liquid oxygen at -219°C has a susceptibility of 310×10^{-6} .

The most strongly diamagnetic element is bismuth, whose susceptibility at 18°C is -1.35×10^{-6} . Next comes antimony with a value of -0.87×10^{-6} . Pure copper is feebly diamagnetic with $\kappa = 0.086 \times 10^{-6}$.

Both the susceptibility and permeability of ferromagnetic bodies vary with the flux density and consequently with the field intensity. But they each have a maximum value. The values of maximum susceptibility range from 1200 for annealed cast steel down to 24 for nickel and 14 for cobalt.

698. Electromagnets. A solenoid without an iron core behaves like a magnet. This may be demonstrated by the attractions and repulsions between two helical coils carrying currents, or between a bar magnet and a pivoted solenoid, as shown in Fig. 100. Here the contacts to the coil are made through mercury cups, and the solenoid is thus free to rotate about a vertical axis passing through its center. When a current is sent through the coil, it tends to point north and

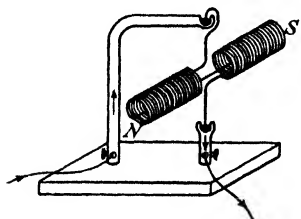


Fig. 100.

south, like a magnetized needle, and one of the poles of a bar magnet repels one end and attracts the other.

In general, electromagnets have iron cores. This makes them much stronger, because the high permeability of the iron produces a flux density numerically much greater than the intensity of the magnetizing field. This flux issuing from the poles of the magnet may be regarded as the external field of force, because in air $B = H$; so the magnetic force on another magnet or on the induced poles in unmagnetized iron is greatly increased.

The magnetizing field H , as we have seen, is weakened by the demagnetizing effect of the poles of the iron core of a straight solenoid, so that equation (2) of Article 695 applies approximately only to a

long solenoid. But this equation is much more exact when the coil is wound in the form of an anchor ring, or **toroid**, as shown in Fig. 101. Then if the mean circumference l , (πD), is large compared to the section area A , the formula applies, for there are no free poles to be considered. Then having found H , we may calculate B if the permeability is known.

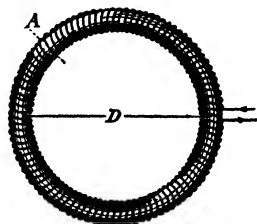


Fig. 101.

Fortunately most practical problems in electromagnetism are similar in principle to that of the anchor ring. In the case of the horseshoe magnet and armature, shown in Fig. 102, if the air gaps between the poles and the bar are short, the condition is near enough to that of the closed ring to

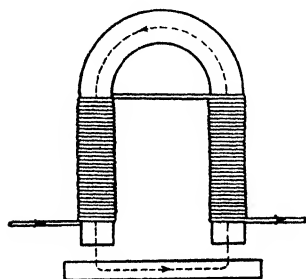


Fig. 102.

permit us to use the same formula in calculating H , and consequently B .

699. Tractive force of a magnet. The force with which an electromagnet holds an armature is commonly called its "lifting power." It would be more logical to call it *holding power*, for a magnet may *hold* a very large weight which it could not *lift* through one millimeter. In the ideal case where the armature makes perfect contact with the magnet

so that there is no leakage flux between the poles, the theory shows that the force required to pull the armature away from the magnet is given approximately by

$$F = \frac{B^2 A}{8\pi}, \quad (1)$$

where B is the flux density, and A is the total area of contact between the two poles and the armature. As B equals the *total flux* divided by the area, $A/2$, of each pole, $B = 2\phi/A$, and we may transform (1) to read

$$F = \frac{\phi^2}{2\pi A}. \quad (2)$$

This means that for the *same total flux*, the tractive force is greatest when the area is least, a conclusion of considerable practical importance in engineering.

The tractive force of a magnet is used commercially for holding large castings or armor plate, as shown in Fig. 103. The magnet is

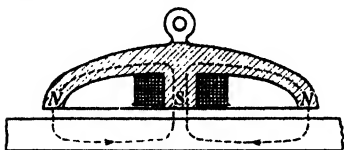


Fig. 103.

is circular in form, like an inverted saucer. There is a resultant south pole at the center, on which the coil is wound, and a continuous ring of north polarity at its edge. This is lowered by a hoisting crane over the object to be lifted; then when

the contact has been made at a convenient point, the current is turned on, and the metal is held by magnetic attraction. The crane then lifts the object and lowers it again at the place desired, when the circuit is broken, and the iron released. In this case the crane does all the lifting, the magnet acting only as a means of attachment.

700. Applications of the electromagnet. Unlike the permanent magnet, the electromagnet may be magnetized and demagnetized at will. This property is made use of in numerous applications. Of these the telegraph sounder and electric bell are among the most important.

The electric telegraph was really invented by the American physicist Joseph Henry in 1831, but made practicable by Morse in 1837. It consists essentially of an armature *A*

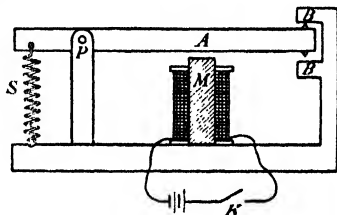


Fig. 104.

(Fig. 104) pivoted at *P*. It is attracted by the electromagnet *M* when the key *K* closes the circuit. A spring *S* holds the armature away from the magnet on open circuit, and as the operator at the sending station manipulates the key, the armature at the receiving end is successively pulled down by the magnet or up by the spring, making a sharp click in each case on the metal arms at *B*. The interval between the down and up strokes may be varied, thus forming what are known as "dots" for short intervals, and "dashes" for long ones. Combinations of these form the letters of the alphabet.

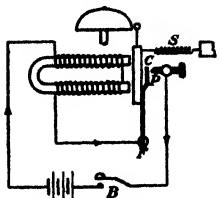


Fig. 105.

The electric bell operates by causing the armature of an electromagnet to vibrate automatically. The armature opens its own circuit when attracted to the magnet, and closes it again when pulled back by a spring. The essential parts are shown in Fig. 105, where the

armature is pivoted at F , and a spring S holds it away from the magnet until the button B is pressed. This results in pulling it toward the magnet, thus separating the flexible strip C from the point P , and so opening the circuit. The armature is then pulled back by S , closes the circuit once more, and the process is repeated. This causes the clapper to strike the bell periodically, as long as the button is held down.

701. Magnetization curves. Permeability bears a relation to magnetic flux density similar to that which conductivity bears

to current density (equation (2), Article 628). But conductivity, as we have seen, is wholly independent of the current, and is a constant at constant temperature. Permeability, however, depends upon

the flux density according to an unknown function of B . This relation can be expressed only graphically, and is usually given indirectly, with B plotted against H , as shown in the accompanying diagram (Fig. 106) for various sorts of iron. But since $\mu = B/H$, we may also plot curves with μ as a function of either B or H , by using the ratios obtained from the B - H curve. Such curves for soft annealed and wrought iron, giving μ as a function of B , are shown in Fig. 107, where μ is seen to increase for small values of B to a maximum, and then decrease as the flux density increases. At 20,000 gauss

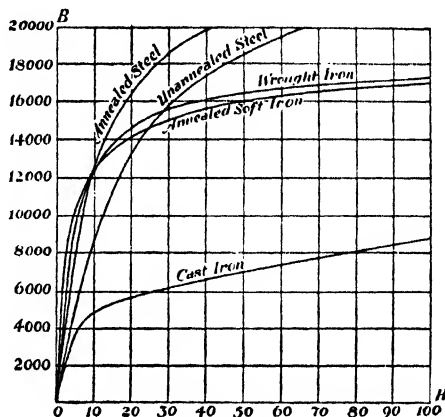


Fig. 106.

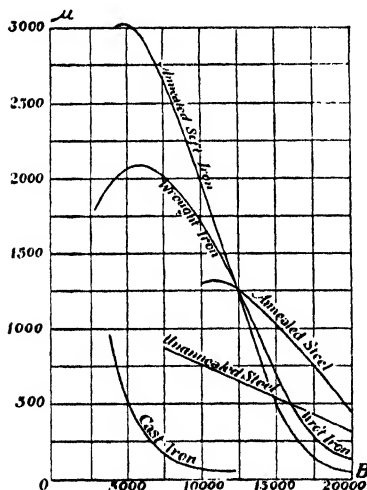


Fig. 107.

the permeability of even the best iron is so small that it does not pay to attempt to force the magnet to a higher degree of saturation. In

commercial machines it is rarely forced much beyond the knee of the B - H curves shown above. Thus 14,000 gauss would be the practical limit for soft annealed iron, whose knee, or point of maximum curvature, is reached when H is about 20 oersteds.

The particular "permalloy" mentioned in Article 566 has a maximum permeability of 105,000, with a flux density of 5000 gauss. This calls for a field of only 0.05 oersted, and a bar of this material may be magnetized to saturation by holding it in line with the earth's field with an intensity of, say, 0.5 oersted. In contrast to permalloy we may cite ordinary wrought iron. Its maximum permeability rarely exceeds 2000, and it requires a field of 2.5 oersteds to develop a flux density of 5000 gauss, and about 50 oersteds to saturate it.

As the permeability of permalloy falls rapidly when B exceeds 5000 gauss, it cannot be used when a large flux density is needed. In order to meet this requirement, another alloy, "permendur," (half iron, half cobalt) has been developed. It has a high permeability which endures into the region of flux densities between 12,000 and 23,000 gauss, and does not reach its maximum permeability of 8000 until B reaches 12,000 gauss.

702. Hysteresis. If a copper wire is clamped at one end, and held vertically, a force F applied horizontally, as shown in Fig. 108 (a), will bend it through a certain distance x . As copper is not highly

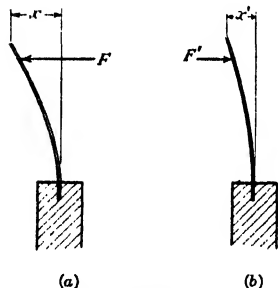


Fig. 108.

elastic, it is easy to exceed the elastic limit of bending. Then if bent beyond the elastic limit, when F is removed, the wire does not return to its original position, but the free end is still displaced a distance x' , as shown in (b). To straighten it, a smaller force F'' must be applied in the opposite direction, and the wire may be held erect under its action. But if the force F' is increased to a particular value F'' , the wire is bent to the right, and if now released, will spring back into

its original vertical position. Thus we see that a certain succession of deviations x corresponds to a certain set of values of F on the outward swing, while a different set corresponds to diminishing F , and increasing F' on the return. This is because the wire is not perfectly elastic within the range of motion considered. As a result it becomes heated when bent vigorously back and forth, because work is done on it in each outward swing, which is not recovered on the return. The

difference, taken over a complete cycle, is the energy consumed in the process, and this appears as heat.

An electromagnet subjected to a varying field H behaves in the manner just described. This behavior may be roughly explained by the lack of freedom of the small dipoles within the iron, as was indicated in Article 565. It amounts to imperfect magnetic elasticity, and is called **hysteresis**, from a Greek word meaning *to lag behind*. Let a sample of unmagnetized iron be subjected to an increasing field, and let the result be plotted on the B - H diagram, as in Fig. 109. The curve will start at O and go to a , like the curves of Fig. 106. But now if H is diminished, the induction B follows a new curve, and when H is zero, there is a residual flux density Ob called **remanence**. The ability of the iron to retain magnetism is called **retentivity**. This makes it possible to produce permanent magnets by raising B to a high value and so obtaining a correspondingly large *residual magnetism* when the magnetizing field is removed. The next portion of the curve, from b to c , is obtained by reversing H , thus reducing B to zero at c , and the magnet may be kept in this state under the action of $H = Oc$. This value of H is called the **coercive force**. If H is now further increased in this reverse sense, a point d , symmetrical with a , may be reached with $-B$ as large as $+B$. A return to a is effected by carrying H back to O , reversing it once more, and building it back to its original maximum value while the curve follows the route $defa$. The cycle of changes is now complete, and may be repeated as often as desired, with B following the same sequence of values, provided H varies between the same limits.

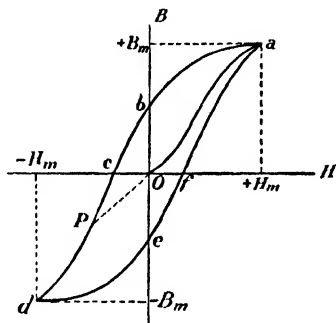


Fig. 109.

It will be seen that the curve never passes through the origin after the start, which means that the iron has not been demagnetized. There must, however, be some point P where, if H were reduced to zero, B would follow the dotted curve to O , and the magnetism would vanish. As P is hard to find, it is easier to demagnetize the iron by causing H to vary less and less from zero, as it goes through successive cycles, until the loop collapses on the origin. This is easily effected by gradually withdrawing a piece of iron from an alternating magnetic field to which it has been subjected.

The area of the hysteresis loop may be proved to be proportional to the work done in carrying the iron through the cycle. It appears as heat, which means wasted energy, and may be very great if the frequency of the cycles is high and the field strong. Hard iron and steel have greater hysteresis losses than soft iron, as shown by the greater width of their loops when plotted between the same maximum values of field strength. Therefore, the core of an electromagnetic machine which carries an alternating or variable flux is made of a grade of iron having as narrow a hysteresis loop as possible. To illustrate the magnitude of the lost energy due to hysteresis, we may take ordinary wrought iron, which consumes about 16,000 ergs per cm^3 per cycle when the magnetization is carried to saturation.

In addition to the alloys mentioned in Article 701, the Bell Telephone Laboratories have developed another called "perminvar" made of 45 per cent nickel, 25 per cent cobalt, and 30 per cent iron. If perminvar is "baked" in the process of manufacture, it acquires the remarkable property of having practically no hysteresis. The B - H curve, if not forced beyond 800 gauss, retraces its path with decreasing field, so as to enclose no area when the material is put through such a cycle. If H varies within the limits of $+2.6$ and -2.6 oersteds, there is no hysteresis and no lost energy.

In contrast to perminvar, an alloy known as "alnico" has been produced by the General Electric Company. This alloy, as its name suggests, is made of aluminum, nickel, and cobalt, alloyed with iron as the main constituent. Its hysteresis curve shows an unusually high *remanence*, which adapts it admirably for use in making permanent magnets. In fact, magnets made of this substance are said to be capable of holding sixty times their own weight.

703. Magnetomotive force. In designing electrical machinery, engineers have to calculate the magnetic flux produced by the field of force developed within a solenoid. This calculation is greatly simplified by the use of a quantity known as **magnetomotive force** (m.m.f.). It is analogous to electromotive force, and is not a force, but measures difference of magnetic potential. It is therefore found in terms of work per unit pole, just as electromotive force is measured in terms of work per unit quantity. Let us apply this principle to the simplest case, that of a solenoid wound in a closed ring, as specified in Article 698. We may imagine that a pole of strength m is carried once around the axis of the ring and against the field H set up by a current in the solenoid. The force on this pole is Hm , and if l is the length of the path (mean circumference), the work done is Hml .

Substituting the value of H given by equation (2), Article 695, we obtain

$$W = 0.4\pi NI m.$$

Then as the m.m.f. is W/m , by definition

$$\text{m.m.f.} = 0.4\pi NI, \quad (1)$$

where I is the current in amperes.

This definition may be regarded as the difference in magnetic level created by NI *ampere turns*. Its c.m.u. is the **gilbert**, named for William Gilbert (Article 555), and one ampere turn creates a m.m.f. of $4\pi/10$ gilbert. The same formula may also be used to calculate approximately the m.m.f. produced by a long slender solenoid, or by a solenoid wound on a straight iron core that forms part of any continuous, or nearly continuous, iron path, as explained in the next article.

704. The magnetic circuit. As was noted in Article 701, B corresponds to current density in a wire carrying a current, and permeability corresponds to conductivity. Following out this analogy with the aid of equation (2) of Article 628 (that is, $I/A = k(dV/dl)$), we see that current density equals the product of conductivity and potential gradient. In magnetism this gradient is the field strength H , just as in electrostatics, field strength equals the space rate of change of potential (equation 2, Article 588). Thus the relation $B = \mu H$ is an exact analogue of the flow of electricity along a conductor.

If we compare equation (1) of Article 703 with (2) of Article 695 (that is, m.m.f. = $0.4\pi NI$, and $H = 0.4\pi NI/l$), it is evident that

$$\text{m.m.f.} = Hl. \quad (1)$$

This is also true for an electric circuit, in which the potential difference between two points may be found by multiplying the potential gradient by the distance between them.

In order to calculate the total flux ϕ , we multiply its density, B , by the section area, A , so that

$$\phi = BA = \mu HA. \quad (2)$$

But $H = 0.4\pi NI/l$; hence

$$\phi = 4\pi NI\mu A/10l. \quad (3)$$

In this expression, $4\pi NI/10 = \text{m.m.f.}$, and denoting m.m.f. by the letter G ,

$$\phi = \frac{\mu AG}{l},$$

or

$$\phi = \frac{G}{l/\mu A} = \frac{G}{Z}. \quad (4)$$

As ϕ is the magnetic equivalent of current, and m.m.f. is the equivalent of e.m.f., the quantity $l/\mu A$, represented by Z , must be the magnetic equivalent of resistance. It is known as the **reluctance** of the magnetic circuit, and when μ is known, it may be calculated from the dimensions of the iron path, just as the resistance of a conductor follows from $R = \rho l/A$ if the resistivity ρ and dimensions of the wire are known. In fact we may make the two expressions identical in form by writing $Z = \sigma l/A$, where σ equals $1/\mu$ and is known as the **reluctivity** of the material. It is now clear that equation (4), $\phi = G/Z$, is the magnetic analogue of Ohm's law, $I = E/R$.

The definitions of the various magnetic and equivalent electrical quantities we have been discussing may be summarized as follows:

	Electric	Magnetic
Specific	Resistivity = ρ Conductivity = $1/\rho = k$	Reluctivity = σ Permeability = $1/\sigma = \mu$
General	Resistance = $R = \rho l/A$ Conductance = $1/R = kA/l$	Reluctance = $Z = \sigma l/A$ Permeance = $1/Z = \mu A/l$

705. Magnetic-circuit problems. In general, the magnetic circuit has a cross section which varies along the path of the flux, and it is often composed of two or more substances; therefore we cannot calculate the flux by a simple substitution in the formula. In practice such problems usually specify the amount of flux that is to be developed in a given circuit, and it is desired to find the value of NI , or the "ampere turns" required to produce it. Then with NI known, the choice of either variable, N for instance, may be made to suit other imposed conditions. These are the available voltage, dimensions of the magnet, allowable heating, and so forth.

As an illustration, suppose the required NI is found to be 12,000 ampere turns. Then I may be 6 amperes with 2000 turns, or 20 amperes with 600 turns, or any combination which gives the necessary total.

To find NI with an assumed ϕ , the values of A and l are estimated for the various materials of the circuit. In Fig. 110 the magnet is supposed to be made of wrought iron having a mean length l_1 . The armature is of soft annealed iron for which the mean length of path of the flux is l_2 . The air gaps, whose perme-

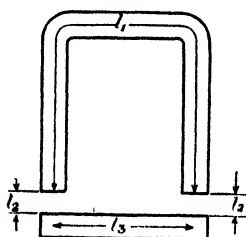


Fig. 110.

ability is unity, have a total length of $2l_2$. The various cross sections A_1 , A_2 , and A_3 are estimated, and the corresponding values of B obtained. Thus (in the magnet core) $B_1 = \phi/A_1$, similarly $B_2 = \phi/A_2$, and $B_3 = \phi/A_3$. We must then find μ_1 and μ_3 from the B - H curves of the two kinds of iron by determining the values of H which correspond to the appropriate value of B . Then $\mu_1 = B_1/H_1$, and $\mu_3 = B_3/H_3$. Finally, remembering that $\mu_2 = 1$, we find that equation (4) of Article 704 becomes

$$\phi = \frac{0.4\pi NI}{\frac{l_1}{\mu_1 A_1} + \frac{l_2}{A_2} + \frac{l_3}{\mu_3 A_3}}. \quad (1)$$

This gives NI in terms of known quantities. The reverse problem of finding ϕ , when NI is given, is much more difficult, and in some cases can be worked only by the method of "trial and error." We assume a value of ϕ , and find μ_1 and μ_3 as before. These values are substituted in equation (1), from which NI is calculated. This result must be made to agree with the actual value by altering the assumed flux.

706. The Barnett experiment. Through the evidence of spectroscopy, it has been found that the electron, rather than the atom, is the ultimate magnet or dipole. Electrons, in addition to "orbital" motion about the nucleus of an atom, have a *spin* like that of a "curved" baseball set spinning by the pitcher. This has two effects. One is purely electromagnetic, producing a magnetic field along the axis of spin, due to the rotating charge, somewhat as in Rowland's experiment (Article 620). The other effect is a gyroscopic action, like that described in Article 113.

An experiment performed by Barnett† in 1914 gives striking evidence of the gyroscopic action. He produced this effect by abruptly magnetizing an iron cylinder suspended vertically, as shown in Fig. 111, where the field H is directed vertically downward. This causes a twist in the axes of the electron "gyroscopes" which tends to make them more nearly parallel with the field. But this twist involves a rotation about an axis normal to their axes of spin. The resultant of the two rotations is a rotation about a third axis, like P in Fig. 79 of Article 113. This experiment proves that the process of magnetizing a bar of iron consists in bringing the axes of electron spin into partial alignment with the field.

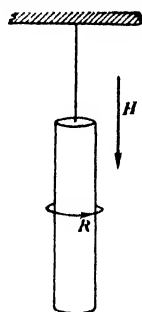


Fig. 111.

† S. J. Barnett, *Physical Review*, 1915, Vol. 6, pp. 171, 239.

The fact that some metals are ferromagnetic and others are not is explained as follows: According to evidence obtained with the spectro-scope, the atoms of nonmagnetic substances have as many electrons spinning in one sense as in the other sense. Thus they are magnetically neutral. But in an iron atom there is an excess of four electrons spinning one way over those spinning the other. In cobalt there is an excess of three, and in nickel, one.

707. The Barkhausen noises. The theory of magnetization outlined above further postulates "domains" within which all the atoms of a ferromagnetic element like iron act upon each other so as to line up with the axes of their spinning electrons parallel to each other. These domains are supposed to be always magnetically saturated, but in the absence of an external field the axial directions of the various groups are oriented wholly at random, so they neutralize each other in the metal as a whole. In an experiment due to Barkhausen, the existence of these domains may be proved by very gradually applying a magnetic field to a sample of iron. This is surrounded by a coil in which currents are induced by any change in the magnetic flux, as is explained in the next chapter. The coil is connected to an amplifier and so to a telephone receiver. While the field is increasing, a succession of clicks is heard, each click corresponding to a turnover of one of the domains. Thus the B - H curve, if sufficiently magnified, would not appear smooth, as in Fig. 106, but like an irregular flight of stairs. Each of the steps would indicate the alignment of a single domain. Careful measurements of the curve and the frequency of the clicks show that the domains which account for them must contain something like 10^{15} atoms and occupy a volume equivalent to a cube having an edge of about 0.025 mm.

SUPPLEMENTARY READING

- C. A. Culver, *Electricity and Magnetism* (Chap. 16), Macmillan, 1930.
A. W. Hirst, *Electricity and Magnetism* (Chap. 12), Prentice-Hall, 1937.
A. Zeleny, *Elements of Electricity* (Chap. 20), McGraw-Hill, 1930.
W. H. Timbie, *Elements of Electricity* (Chap. 6), Wiley, 1925.

PROBLEMS

1. Calculate the exact value of the field at the center of a solenoid of 200 turns, whose length of 20 cm is not great compared to its radius of 5 cm, when a current of 4 amperes flows through it. Also find H , assuming r negligible compared to l . *Ans.* 44.94 oersteds; 50.24 oersteds.

2. A ring of soft annealed iron is wound with 400 turns of wire. The ring's mean radius is 18 cm, and that of its section 1.5 cm. Using the curve of Fig. 106, calculate the current required to send 113,000 maxwells through the core. *Ans.* 11.2 amperes.

3. If the ring in Problem 2 is broken with an air gap of 2 mm width, calculate the current required to produce the same flux.' *Ans.* 17.6 amperes.

4. An electromagnet and the armature that bridges its poles are made of annealed steel and have an average section of 4 cm², and an average length of magnetic path of 24 cm. The magnet is wound with 172 turns through which flows a current of 3 amperes. Calculate the flux and lifting power of the two poles. *Ans.* 72,000 maxwells; 103.2 megadynes.

*5. It is desired to send 400,000 lines of induction through the armature of a generator specified as follows: The field core of cast iron has an average section of 64 cm² and length of 90 cm. The path of the flux through the armature of soft annealed iron is 35 cm long, and has an average section of 25 cm². The air gaps between pole pieces and armature have an effective section of 80 cm², and are each 3 mm across. Using Fig. 106, calculate the ampere turns required. *Ans.* 5950 ampere turns.

CHAPTER 52

Induced Currents

708. Induced electromotive force. By far the most important method of producing an e.m.f. is electromagnetic. Without the intervention of chemical, thermal, or radiant energy, we may transform mechanical into electrical energy at almost any desired voltage. This important fact was discovered by Faraday in 1831. The transformation consists essentially of relative motion between a magnet and a conductor, resulting in a current through the conductor. The

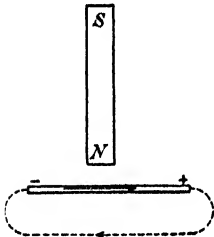


Fig. 112.

discovery was as epoch-making as Oersted's discovery eleven years earlier. Oersted found that an electric current could move a magnet. Faraday discovered that a moving magnet could produce an electric current. One effect is the inverse of the other, but both are fundamental facts in the theory of electromagnetism, and had to be discovered independently.

Let a metal rod be moved at right angles to its axis across the pole of a magnet, so that it cuts the flux emanating from the pole. Then the potential of one end is higher than that of the other, as long as the rod is moving. Thus in Fig. 112, if the conductor moves perpendicularly to the plane of the paper and toward the observer, the potential of the right-hand end is higher than that of the left-hand end. If the ends are connected by a wire, indicated by the dotted line, a current will flow from low to high potential through the rod, and from high to low in the wire. Therefore the moving "inductor" is the seat of an electromotive force directed from left to right.

709. Lenz's law. Let the operation just described be viewed along the axis of the rod, whose section is shown by the central circle in Fig. 113. Then the current, when the circuit is completed, is flowing away from the observer. The lines of force which surround it are directed as shown by the dotted arrows, and the direction of motion is shown by the solid arrow. The direction of the lines of

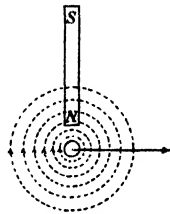


Fig. 113.

force is the one in which a north pole would move in the field that the lines represent. It is then evident from the diagram that the north pole of the inducing magnet is urged to the right, and thus tends to keep up with the moving rod. If it is prevented from moving, this force acts as a drag on the rod and must be overcome while it is in motion. Thus work must be done upon the moving rod equal to the product of the magnetic drag times the distance moved.

The current obtained by the above process is produced at the expense of mechanical energy. The current represents energy, and we find that it cannot be induced without the expenditure of an equivalent amount of work elsewhere. This sounds very much like the law of the conservation of energy. Further, we may affirm that *the direction of the induced current must be such as to create a field opposing the relative motion that caused it.* This is Lenz's law, which is really a statement of Newton's law of action and reaction extended to cover electromagnetic forces. It should be noted that on *open* circuit, as no current flows, no appreciable amount of work is done in creating merely a difference of potential.

710. Direction of the induced current. This can always be obtained from the known direction of the flux surrounding the current, combined with Lenz's law. But it is more convenient to use the "right-hand rule" as follows: Extend the thumb, index, and middle fingers so as to approach mutually perpendicular directions, as shown in Fig. 114. Then if the thumb represents the direction of the motion of the inductor, and the index finger the direction of the flux, the middle finger points in the direction of the current. It is easy to remember this rule by recalling the fact that the most mobile finger, the thumb, stands for motion. This is one of the two causes of the induction.



Fig. 114.

The index finger stands for the second cause, or flux, while the third represents the direction of the effect, or current. We may, however, choose any finger to represent motion, provided we follow the same order: motion, flux, current, in the same sequence of the fingers.

711. Essential conditions of induction. To induce a current in a conductor there must be relative motion between magnet and conductor. But Faraday saw that because of this relative motion, the magnetic flux was cut by the conductor. In fact he created the fiction of a magnetic flux in order to form a useful picture of the process of inducing a current. We may then say that an e.m.f. is induced whenever a conductor moves across a magnetic flux, or whenever a

magnetic flux moves across a conductor. It makes no difference which moves, provided the flux is cut, and we may generalize by saying that *an induced e.m.f. is caused by a cutting of the flux*. This may be due to relative motion between a magnet and a conductor, but it is also possible to cut the flux when there is no motion, as will be explained. Then **flux** may be defined as *the condition of space in and around a magnet which gives rise to an induced e.m.f. in a conductor when relative motion of flux and conductor results in cutting the flux*.

We must now qualify the right-hand rule for finding the relative directions of motion, flux, and current. The thumb points in the direction of the *relative motion of the conductor* (or *inductor*, as it may be called). Therefore if the magnet in Fig. 113 is moved instead of the inductor, the thumb points in a direction opposite to the motion of the magnet.

If a bar magnet is moved along the axis of a coil of wire, as shown in Fig. 115, the flux from the north pole cuts the turns of the coil,

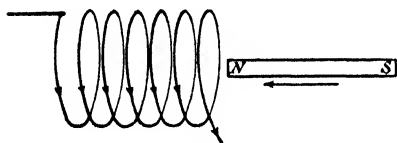


Fig. 115.

and an e.m.f. is induced. According to Lenz's law, this tends to send a current around the coil in such a direction as to create a field in opposition to the motion of the pole. In this case it is clockwise when

viewed from the left face of the coil, looking toward the advancing magnet, in accordance with the usual right-handed screw rule. Further, the more rapidly the magnet is moved, the greater the induced e.m.f. and the larger the resulting current.

If a coil *A* carrying a current is substituted for the magnet, as in Fig. 116, the same effect is produced. When its current flows so as to create north polarity on the face which is approaching

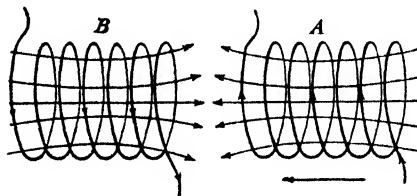


Fig. 116.

coil *B*, then motion of *A* toward *B* induces a current in *B* which flows as in the preceding case. In both the case of moving magnet and moving coil, a reversal either of polarity or direction of motion reverses the direction of the induced current.

The energy of the induced current, in the case illustrated by Fig. 115, is created wholly at the expense of the mechanical work needed to push the magnet into the coil or to pull it out again. It is not de-

rived from the magnet. But in the case shown in Fig. 116, the energy is supplied in part by mechanical work, and in part by the current in coil *A*. This is because the flux produced by the current in *B* reacts upon *A*, and in overcoming this reaction the current in *A* supplies energy.

712. Induction without motion. This case cannot be said to be in quite the same class as the sources of e.m.f. we have just been discussing, where one form of energy, that is, mechanical, was used at least in part to create another form, that is, electrical energy. In order to induce currents without motion and with no expenditure of mechanical energy, there can be a transfer only of electrical energy from one coil to another.

If a current is *started* in coil *A*, Fig. 116, and if the coil is kept at rest, a current is induced in *B* exactly as if *A* were moved toward it. The flux sent out from *A*, when its circuit is first closed, cuts the coil *B* in the same manner as when a steady current was flowing in *A* as it moved toward *B*. But here the rate of increase of the current in *A* determines the magnitude of the induced e.m.f. None of the energy of the induced current is now derived from mechanical work, but wholly from electrical energy consumed in *A* (the *primary* coil) while the current is being established. This exceeds the amount when no current is induced in *B* (open circuited *secondary*) by the exact value of the energy developed in the secondary on closed circuit. Thus it appears that the law of the conservation of energy may be extended to include electromagnetic phenomena.

As soon as the current in the primary has been established at its normal value determined by Ohm's law, no more change in the flux occurs, and the transfer of energy to the secondary ceases. But now if the current is broken, the flux through the primary collapses and produces the same effect on the secondary as when the coils were separated with a steady current flowing in one of them.

In all cases of induced currents, one fundamental principle is the cause of the induction, and that is the cutting of a magnetic flux by the inductor. As the rate of cutting determines the magnitude of the induced e.m.f., we may state the laws of induction thus: *Induced electromotive force varies directly as the time rate of cutting of the magnetic flux, and tends to cause a current to flow in a direction which would set up a flux in opposition to the motion, or change of flux, which caused it.*

713. Magnitude of the induced e.m.f. We shall now derive relations enabling us to calculate the value of the e.m.f. as determined by the rate at which an inductor cuts magnetic flux.

As was proved in Article 618, the field *H* at a point *r* centimeters

from a long wire carrying a current is given by $H = 2I/r$. Therefore in Fig. 117 the force F on a pole m at the point p is $2Im/r$. Now suppose

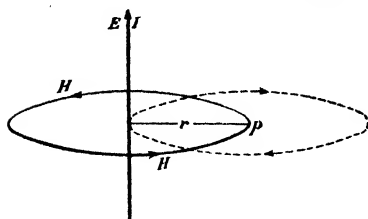


Fig. 117.

the pole to be carried at a constant rate once around the wire in opposition to H , and in a circle of radius r , or suppose that the wire is carried once around the pole at the same distance, in a path shown by the dotted line. Then in both cases, which are really the same, every line of force emanating

from the pole is cut once. The work done is $2\pi rF$; or substituting for F , we have

$$\begin{aligned} W &= 2\pi r \times \frac{2Im}{r} \\ &= 4\pi Im. \end{aligned} \quad (1)$$

But this action causes an equal amount of work to be done in the wire, which appears as the electrical energy $-EQ$, where Q is the quantity of electricity which passed through the wire during the time t of the action, and $-E$ is an induced e.m.f. This e.m.f. is so directed as to oppose the motion which produced it, as indicated by the minus sign. But as the motion was supposed to be against the field, the e.m.f. favors the current already flowing, as indicated in the diagram.

We may now calculate E by setting the mechanical work done equal to the electrical energy produced. Then

$$-EQ = 4\pi Im.$$

But

$$Q = It.$$

$$\therefore EIt = -4\pi Im,$$

and

$$E = -\frac{4\pi m}{t}. \quad (2)$$

Since $4\pi m$ is the flux ϕ emanating from the pole m , we may write $E = -\phi/t$. Therefore if the rate at which the flux is cut is constant, the induced e.m.f. is equal to the total flux divided by the time it takes to cut it. If the rate varies, we must set E equal to $-d\phi/dt$, which is the instantaneous value of the e.m.f. whether the flux or the velocity or both are constant or not.

In the preceding discussion we have supposed that a current was already flowing in the wire. But as this current appears on both sides of the equation of energies, it cancels out, and the final result is inde-

pendent of its value. Therefore equation (2) applies even when no current is flowing. Thus an e.m.f. is induced even when the wire is not part of a closed circuit. If it is part of a closed circuit, an *induced current* is set up in that direction which opposes the relative motion of the pole and the wire.

Since neither r nor m enters specifically into the basic expression $E = -d\phi/dt$, we may infer that when the flux which passes through a coil of N turns is varied by any method, we may write

$$E = -Nd\phi/dt. \quad (3)$$

Experiment shows that this inference is correct both when the coil moves with respect to the flux, and when the coil itself remains stationary and the change in flux is due solely to an alteration in the strength of the magnetic field.

Now suppose the flux through a coil of N turns increases from an initial value ϕ_1 , to a final value ϕ_2 ; then if the rate is uniform

$$E = -\frac{N(\phi_2 - \phi_1)}{t}. \quad (4)$$

As this gives E in e.m.u., we must divide by 10^8 to reduce it to volts, as explained in Article 626. Then the general expression, which applies either to a constant or to a variable rate of change of the flux, becomes

$$E \text{ (volts)} = -\frac{N}{10^8} \frac{d\phi}{dt}. \quad (5)$$

714. Magnitude of the induced current and quantity. The important result given above enables us to calculate the instantaneous induced e.m.f. whenever there is relative motion of flux and inductor. When the inductor or coil is part of a closed circuit, a current must flow through it as a result of the induced e.m.f. Then, applying Ohm's law, we obtain

$$I = \frac{E}{R} = \frac{-N}{10^8 R} \frac{d\phi}{dt}, \quad (1)$$

which is the instantaneous value of the induced current. Transposing, we have

$$Idt = -N \frac{d\phi}{10^8 R} \quad (2)$$

where Idt is the infinitesimal quantity dQ which flows in the time dt . If the time has a finite value t , the current varies during that time, and as we do not know its average value, It cannot be calculated directly. However, the total quantity Q may be found from equation (2) as follows: If ϕ changes from an initial value ϕ_1 to a final value

ϕ_2 , the sum of the infinitesimal changes in ϕ is $\phi_1 - \phi_2$, and equation (2) becomes

$$Q = +N \frac{\phi_2 - \phi_1}{10^8 R}. \quad (3)$$

This means that the quantity of electricity in coulombs which passes through the circuit when the flux changes from ϕ_1 to ϕ_2 is independent of the way in which the flux was cut, or of the varying rate of cutting it, and depends only upon the initial and final conditions.

715. The ballistic galvanometer. Equation (3) of the last article is of great value in the comparison of capacitances or other apparatus in which the discharge occurs suddenly, and a certain quantity of electricity flows through the circuit in a very short time. This is most readily done by using a ballistic galvanometer, which is merely a moving-coil instrument whose armature has sufficient inertia to be rather slow in starting to swing. Its angular momentum then carries it through an angle proportional to S , the initial impulse. The impulse is equal to the torque L applied to the armature, multiplied by the time during which the impulse acted, or in general, $dS = Ldt$. But the torque varies as the current; therefore $dS \propto Idt$, where Idt is the quantity dQ which flows in the time dt . We may then affirm that the impulse S is proportional to the total quantity of electricity that flows through the galvanometer. If the discharge is sufficiently rapid, the resulting "throw" of the armature (galvanometer coil) is proportional to S and, therefore, to Q .

If we use the principle just established, the ballistic galvanometer enables us to compare the capacitance of an unknown condenser C_x with that of a standard condenser, C_s . These are charged by the same battery to the same potential ΔV , and then discharged through the galvanometer. Since $C = Q/\Delta V$, the charge on each condenser is directly proportional to its capacitance, or $C_x/C_s = Q_x/Q_s$. But the deflections d_x and d_s are proportional to the quantities discharged, or $d_x/d_s = Q_x/Q_s$. Therefore

$$\frac{C_x}{C_s} = \frac{d_x}{d_s},$$

from which C_x may be found in terms of known values.

716. Mutual induction. The method of inducing currents without motion, which was explained in Article 712, is described as **mutual induction**, because each coil acts upon the other as soon as a current flows in the so-called "secondary" coil, as has already been noted. This arrangement is shown in Fig. 118. Mutual induction occurs when the current in the primary is varied so as to act inductively

upon the secondary by means of the mutual interlinkage of part of the primary flux. Since the amount of interlinking of this flux with the secondary coil depends upon the position and dimensions of both coils, and the medium between them, the induced e.m.f. cannot be calculated except in some very simple cases.

In general, we must know from previous measurements a constant of the system called the *coefficient of mutual induction*, or the **mutual inductance**. This is defined as the ratio of the induced secondary e.m.f. to the time rate of change of current in the primary, and is expressed by

$$E_2 = -M \frac{dI_1}{dt}, \quad (1)$$

where M is the mutual inductance. Now $E_2 = -N_2 d\phi'/dt$, where ϕ' is that portion of the primary flux which interlinks the secondary, and N_2 is the number of turns of the secondary coil. The negative sign means that when the current I_1 is increasing (dI_1/dt positive), E_2 tends to send a current in the opposite direction through the secondary, and thus builds up an opposing flux which resists the growth of the interlinking flux. When the current is decreasing (dI_1/dt negative), E_2 tends to send a current through the secondary

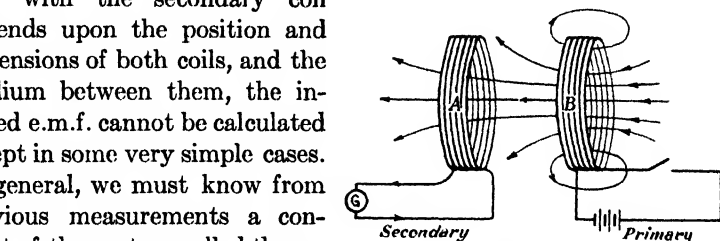


Fig. 118.

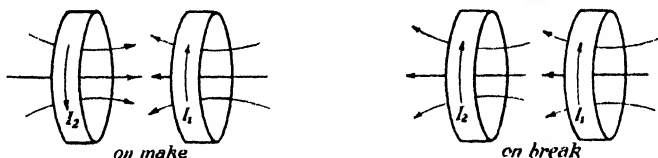


Fig. 119.

in the same direction, and thus retards the decay of ϕ' . These effects are suggested in Fig. 119. We may now substitute $E_2 =$

$-N_2 \frac{d\phi'}{dt}$ in the equation which defines M , and obtain

$$N_2 \frac{d\phi'}{dt} = M \frac{dI_1}{dt},$$

or

$$N_2 d\phi' = M dI_1.$$

But since the flux is zero when the current is zero, and is equal to ϕ' when the current is I_1 , we may use the same kind of reasoning as

in Article 714, and state that when the final current has fully built up the final flux,

$$N_2\phi' = MI_1.$$

$$\therefore M = \frac{N_2\phi'}{I_1} \text{ e.m.u.,}$$

and when I_1 is measured in amperes,

$$M = \frac{10 N_2\phi'}{I_1} \text{ e.m.u.} \quad (2)$$

These last expressions enable us to describe M as the total amount of secondary interlinkage ($N_2\phi'$) per unit primary current, since every line of the ϕ' flux produced by I_1 links with every one of the N_2 turns of the secondary. The coefficient M may now be calculated provided we can calculate ϕ' for a given current. This is possible approximately for the simple case when the secondary is wound closely over the center of a long solenoid serving as primary (Fig. 120),



Fig. 120.

or on a closed ring, if we assume that both are wound on nonmagnetic cores. In these cases the flux due to I_1 flowing in the primary coil S_1 is equal to $4\pi N_1 I_1 A / 10l$, as given in equation (3), Article 704, where A is the coil's sectional area, l is its length, and $\mu = 1$. This is practically identical with the flux that interlinks the N_2 turns of the secondary S_2 . Therefore, substituting for ϕ' in equation (2) above, we have

$$M = 4\pi N_1 N_2 A / l \text{ e.m.u.} \quad (3)$$

The practical unit of mutual inductance is the **henry**, named for the American physicist Joseph Henry (1797–1878) of Princeton. It is defined as *the amount of inductance which will result in an induced e.m.f. of one volt when the current varies at the rate of one ampere per second*. Its dimensions are those of $E/(I/T)$, which is the same as resistance multiplied by time, or $[LT^{-1} \times T] = [L]$. So the unit was called an *ohm-second* before the name *henry* was adopted. It might also be expressed as a length. As the ohm is 10^9 times the absolute unit (Article 624), the henry is 10^9 times the absolute unit of inductance. Therefore equations (2) and (3) above become

$$M = \frac{N_2\phi'}{I_1 \times 10^8} \text{ henries}$$

and

$$M = \frac{4\pi N_1 N_2 A}{10^9 l} \text{ henries.} \quad (4)$$

In practice the henry is rather too large as a unit, and the millihenry, one thousandth of a henry, is commonly employed.

717. Self-induction. Whenever a current flows around a circuit, there is always some magnetic flux which interlinks with it, as shown in Fig. 121, so that, when the circuit is closed, the flux must be established, and must collapse when the current ceases. This is equivalent to cutting the conducting wire by the total amount of the flux in either case, and this cutting process induces an e.m.f. in the circuit in such a direction as to oppose its cause. Then on the "make," the e.m.f. is directed against the current, tending to retard its growth from zero to full value, and on the "break" it acts with it, tending to prolong the time of decay down to zero again.

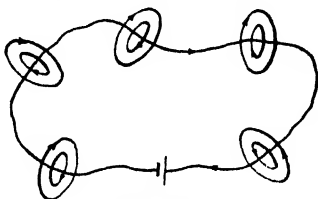


Fig. 121.

This effect can be greatly enhanced by using a coil of many turns of wire through which a strong flux passes when the circuit is closed,

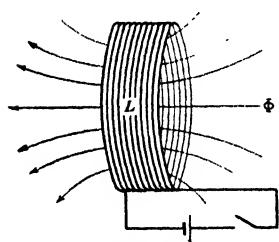


Fig. 122.

as in Fig. 122. The e.m.f. thus induced when the current is growing is called the back electromotive force, or counter-electromotive force, and the whole phenomenon is known as **self-induction**. In general, the flux interlinking a circuit when a known current is flowing can be calculated only for very simple cases such as the long solenoid, or a closed ring. But as in mutual induction, a constant may be defined,

known as the *coefficient of self-induction*, or the *self-inductance*, or simply **inductance**, which enables us to calculate the back e.m.f.

This coefficient is the *ratio of the induced e.m.f. to the time rate of change of the current*; therefore

$$E = -L \frac{dI}{dt} \quad (1)$$

when L is the coefficient, and the negative sign indicates the fact that the e.m.f. is *opposed* to the current when I is increasing (dI/dt positive), and is *with* the current when I is decreasing (dI/dt negative).

As in the case of mutual induction, we may substitute $E = -N(d\phi/dt)$; hence

$$Nd\phi = LdI.$$

This expression, as in mutual induction, may be extended to the form

$$L = N \frac{\phi}{I} \text{ e.m.u.,}$$

or

$$L = \frac{10N\phi}{I} \text{ e.m.u.} \quad (2)$$

when I is measured in amperes. Thus inductance is shown equal to the total amount of interlinkage per unit current. The total interlinkage may be calculated from the known value of ϕ produced by a given current in a closed ring whose core is nonmagnetic. Using, as before, equation (3) of Article 704 and setting $\mu = 1$, we have $\phi = 4\pi NIA/10l$. Substituting this value in (2), we obtain

$$L = \frac{4\pi N^2 A}{l} \text{ e.m.u.,} \quad (3)$$

which shows that L depends upon the dimensions of the coil and the square of the number of turns.

The practical unit of self-inductance is the henry, the same as for mutual inductance, because its defining equation involves the same units in the same manner. Its dimensions are therefore the same, and it must be measured in the same way. Also, as the henry is 10^9 times as large as the e.m.u., equations (2) and (3) become

$$L = \frac{N\phi}{10^9 I} \text{ henries, and } L = \frac{4\pi N^2 A}{10^9 l} \text{ henries.} \quad (4)$$

The effect of self-induction in opposing the growth of a current, and in prolonging its decay, shows that a current with its interlinking flux has inertia exactly as has mass. Both resist change of motion. This fact is brought out strikingly by comparing the expression that gives the energy required to establish a current in a coil with that of the kinetic energy of a moving mass. The coil's energy may be proved to be given by

$$W = \frac{1}{2} LI^2,$$

and the energy of the moving mass is

$$W = \frac{1}{2} mv^2.$$

As kinetic energy is a consequence of the inertia of a moving mass, it is evident that electricity in motion, like ordinary matter, also has inertia, and L is the electrical equivalent of mass.

718. The spark coil. The energy established in a coil in which a current interlinks with a magnetic flux may be utilized to produce a

very high voltage and a hot spark when the current is abruptly broken. In fact most of the stored-up energy may be concentrated in the spark at the point where the current is interrupted. To obtain the best results, and to avoid eddy currents (discussed in Article 721), a coil of many turns should be wound on a core made of a bundle of soft iron wires. The core is not closed upon itself as in a transformer, because, if it were, the residual magnetism would be very large, and so greatly reduce the total change of flux when the circuit is broken. With the arrangement shown in Fig. 123, the instantaneous e.m.f. of self-induction may be roughly indicated by the throw of a voltmeter when the circuit is opened by the switch *K*. It is much larger than the voltage of the battery which magnetizes the core, because the induced e.m.f. depends upon the time rate of change of the flux. This has

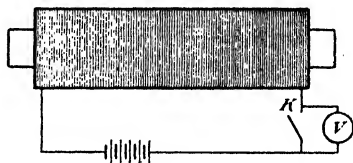


Fig. 123.

nothing to do with the voltage of the battery, and when the magnetic circuit is mostly in air, ϕ falls nearly to zero with great rapidity. With a single dry cell giving 1.5 volts we may develop instantaneously several hundred volts across the terminals of the switch when it is opened. If there is no voltmeter to absorb the induced current, a bright spark due to the "extra current," as it is sometimes called, jumps the gap and may be used as a means of igniting an explosive mixture of gases. This is the basis of the "make and break" method of ignition in certain types of internal combustion engines. We may make this effect even more striking by breaking the circuit under oil, thus reducing the time of interruption of the current; the discharge

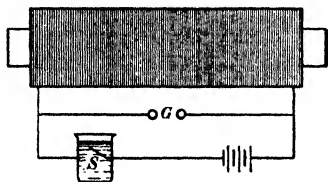


Fig. 124.

may then be made to take place across a small gap *G*, in parallel with the switch *S*, as seen in Fig. 124.

719. The induction coil. If the coil just described is wound with a relatively small number of turns of coarse wire, and this has wound over it many turns of fine wire, the latter coil, or secondary winding, is linked by most of the flux which passes through the primary. It is therefore the seat of an induced e.m.f. whose instantaneous value in volts is

$$E_2 = - \frac{N_2 d\phi}{10^8 dt}.$$

This arrangement, often called the *Ruhmkorff coil*, is named after its inventor.† If N_2 and the total flux ϕ are very large, and if the time rate of establishing or destroying the flux is very short, E_2 may momentarily reach a value of hundreds of thousands of volts and produce a spark many inches long at G , shown in Fig. 125. But as we have

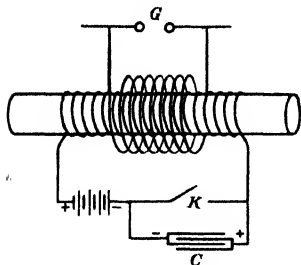


Fig. 125.

seen, it is impossible to establish the primary current, and consequently the flux, with extreme rapidity, owing to the opposing e.m.f. of self-induction. It is therefore necessary to rely upon the break to obtain the high potentials referred to, for the time of decay may be indefinitely shortened. This is greatly accelerated by placing a condenser C across the point of interruption, as shown in the diagram. It absorbs the

“extra current” which would otherwise maintain an arc between the points of contact after the break. As a result, the e.m.f. of self-induction, which tends to keep the current flowing, charges the condenser instead of maintaining the arc. As this induced voltage is far higher than that of the battery, the condenser at once discharges through the primary circuit against the battery’s e.m.f. This reverse current has the desirable result of sweeping out the residual magnetism of the core, and so rendering the total change of flux greater than it would be otherwise. Thus the condenser accomplishes three objects: It prevents the gradual burning up of the contacts, usually of platinum, by the arc at K ; it decreases the time of the break, and it increases the total change of the flux.

In order to make the process of making and breaking the circuit automatic and continuous, various types of interrupters are used in place of a simple switch. Most of them operate on the same principle as the electric bell, and may use the core of the coil itself as the attracting magnet, or may be quite independent of it. The Wehnelt interrupter, however, operates on a very different principle, which is briefly as follows: The primary current passes through an electrolytic cell having dilute sulphuric acid as the electrolyte, as shown in Fig. 126; a strip of lead

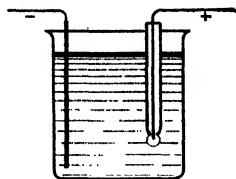


Fig. 126.

† H. D. Ruhmkorff, 1803–1877, a mechanician of German origin who lived in Paris.

is the cathode, and a platinum wire, projecting a short distance below the glass or porcelain tube that surrounds it, is the anode. When the circuit is closed, the current density at the small projecting tip of platinum is so high that it vaporizes the liquid, forming a nonconducting bubble of vapor which envelops the tip and breaks the circuit with great abruptness. Then the high e.m.f. of self-induction throws off the bubble, the current is re-established, and the process is continued.

720. The transformer. As we have seen, the magnetic circuit of an induction coil is not a closed iron path. This fact facilitates a rapid decay of the flux from its maximum value to near zero, and consequently results in a high induced e.m.f. But the flux in the core of a transformer depends upon an impressed alternating e.m.f. which, as we shall see, varies harmonically and not abruptly. The problem then is not to obtain the highest possible induced voltage, but rather to obtain an efficient transformation of electrical energy. This calls for a maximum flux produced with a minimum current and minimum magnetic leakage. It is accomplished by having a continuous iron core wound with both primary and secondary coils, as shown in Fig. 127.

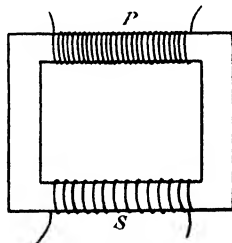


Fig. 127.

The theory of the transformer depends upon assuming an harmonically varying flux interlinking the coils. This induces a counter-e.m.f., E' , in the primary, and a secondary e.m.f., E_2 , in the secondary. As practically all the flux interlinks both coils, these voltages obviously are to each other as the ratio of the number of turns N_1 and N_2 in the coils, or $E'/E_2 = N_1/N_2$.

The primary voltage E_1 must obviously be larger than E' in order to produce the current. But in an unloaded transformer, the excess is very small, and we may use the approximate relation

$$E_1/E_2 = N_1/N_2. \quad (1)$$

This is less exact when the transformer is loaded, but even then it may be used as a first approximation.

Large transformers are remarkably efficient. They have attained efficiencies as high as 99 per cent, so that we may assume 100 per cent in rough calculations. Then power input equals power output, and

$$E_1 I_1 = E_2 I_2. \quad (2)$$

When this is combined with (1), there results

$$I_1/I_2 = N_2/N_1. \quad (3)$$

Thus the ratio of the primary to the secondary current of a loaded transformer equals the inverse ratio of the number of turns in the coils. This means that small currents are associated with high voltages, and vice versa, a fact which indicates the chief purpose of transformers. By "stepping up" the voltage of a generator, we may transmit a given amount of power, using a small current flowing over wires of correspondingly small section, thus saving enormously in the cost of a transmission line. At the receiving end of the line the voltage is

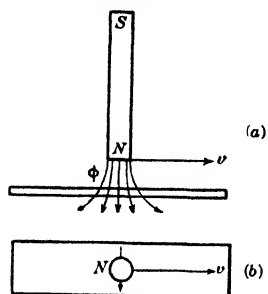


Fig. 128.

"stepped down" again by another transformer to a value suitable for commercial use. In this way the current delivered to the consumer is much greater than that which was carried by the transmission line.

721. Eddy currents. Whenever a magnetic flux passing through a conductor is made to vary either in position or magnitude, it induces an e.m.f. which causes a current to flow, if there is any semblance of a closed circuit. To illustrate this, suppose the north

pole of a bar magnet is drawn across the face of a metal sheet with a velocity v , as indicated in Fig. 128. The vertical flux cuts the inductor from left to right, which means that the inductor's relative motion is from right to left. Then applying the right-hand rule, we find that an e.m.f. is set up at right angles to v and ϕ . This will cause currents to flow in the sheet in the closed circuits or *eddies* shown in Fig. 129, and lying

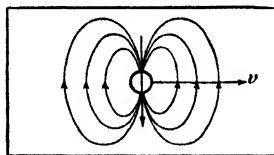


Fig. 129.

both in front of and behind the pole as it moves. The eddy in front of the pole tends to produce a north pole above the inductor sheet, as

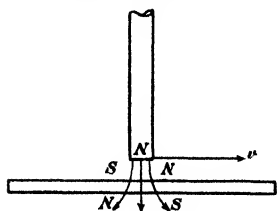


Fig. 130.

indicated in Fig. 130, while the eddy behind the pole tends to produce a south pole above the sheet. Thus the leading eddy is continuously repelled by the moving magnet, and the lagging eddy continuously attracted. This results in a force tending to drag the sheet after the moving pole, or what is the same thing, it opposes the pole's motion. This fact could

have been predicted from the general principle that the electromagnetic reaction always opposes the action which causes it.

Eddy currents may involve a serious waste of energy, for though the e.m.f. induced is usually small, the resistance of the conducting circuit is generally minute, and large currents may flow, with correspondingly large I^2R losses and consequent heating of the inductor. This loss may be greatly reduced by using laminated cores, as indicated in Fig. 131, which shows the section of the armature of a generator. The eddy currents indicated by the curved arrows have to cut across the laminations as well as a thin film of iron oxide which acts as an insulating layer between them. Thus they encounter a great number of relatively high resistances in series. As the e.m.f. induced in the armature core is never large, the resulting eddy currents are reduced to an almost negligible value.

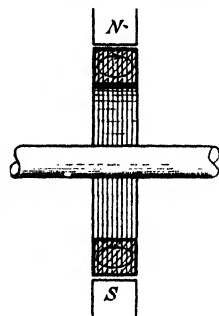


Fig. 131.

722. Relations between units. In this chapter we have completed the list of the essential electromagnetic units. A recapitulation at this point is therefore desirable, as well as a comparison with the same units in the electrostatic system. This comparison will be made easier by reference to the accompanying table of dimensions, most of which have already been derived in the text.

Name	Symbol	c.s.u.	Article	e.m.u.	Article	Ratio
Quantity.....	Q or q	$M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}$	581	$M^{\frac{1}{2}}L^{\frac{1}{2}}$	622	LT^{-1}
Potential, Electro- motive Force....	V or ΔV E or e.m.f. }	$M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}$	587	$M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-2}$	626	$L^{-1}T$
Capacitance.....	C	L	599	$L^{-1}T^2$	692	L^2T^{-2}
Current.....	I or i	$M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-2}$	617	$M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}$	617	LT^{-1}
Resistance.....	R or r	$L^{-1}T$		LT^{-1}	624	$L^{-2}T^2$
Magnetic Pole.....	m			$M^{\frac{1}{2}}L^{\frac{1}{2}}T^{-1}$	557	
Field Strength.....	E and H	$M^{\frac{1}{2}}L^{-\frac{1}{2}}T^{-1}$	585	$M^{\frac{1}{2}}L^{-\frac{1}{2}}T^{-1}$	559	1
Self-induction.....	L }			L	716	
Mutual Induction..	M }					

The last column gives the ratio of the dimensions of a unit as defined in the electrostatic system to the same unit in the electro-

magnetic system, and five of the ratios are seen to be some power of L/T . That is, either a velocity or the square of a velocity appears in each. As was pointed out in Articles 617 and 692, the numerical value of this ratio is found by experiment to be equal to the velocity of light, or approximately 3×10^{10} cm/sec.

SUPPLEMENTARY READING

C. A. Culver, *Electricity and Magnetism* (Chap. 17), Macmillan, 1930.

A. Zeleny, *Elements of Electricity* (Chap. 17), McGraw-Hill, 1930.

PROBLEMS

1. How many volts are generated in a wire 12 cm long which cuts directly across a flux whose intensity is 14,000 gauss, if it moves at a speed of two m per sec.? *Ans.* 0.336 volt.

2. A coil of 150 turns and having a radius of 15 cm is interlinked by a field of 12,000 gauss. This decreases steadily to zero in 6 sec. What is the e.m.f. induced? *Ans.* 2.12 volts.

3. If the resistance of the coil in Problem 2 is 3 ohms, what quantity of electricity has flowed through it at the end of 4 sec., and how many joules of heat energy were evolved? *Ans.* 2.83 coulombs; 6.00 joules.

4. A coil of 80 turns, and having a radius of 15 cm and a resistance of 4 ohms, lies flat on a table where the vertical component of the earth's magnetic field is 0.57 oersted. The coil is connected to a ballistic galvanometer of 12 ohms resistance. It is quickly turned over so as to fall inverted on the table. How many coulombs flow through the circuit? *Ans.* 40.27 microcoulombs.

5. A solenoid produces a flux of 30,000 maxwells in an iron core when a current of 2 amperes flows through it. A secondary coil of 360 turns is wound over it. Assuming that all the flux interlinks both coils, calculate the coefficient of mutual induction. *Ans.* 54 millihenries.

6. A coil of 200 turns is wound on a wooden ring whose section is 12 cm² and whose mean length is 30 cm. Upon this coil is wound another of 100 turns. Calculate the coefficient of mutual induction. *Ans.* 0.1 millihenry.

7. Calculate the coefficient of self-induction of the primary winding in Problem 5 if it has 400 turns, and if the same current is flowing. *Ans.* 60 millihenries.

8. What are the coefficients of self-induction of the primary and secondary coils in Problem 6? *Ans.* 0.2 millihenry; 0.05 millihenry.

9. What is the value of L in Problem 6 if the two coils are connected in series so that their fields are added? What if the connections are reversed? *Ans.* 0.45 millihenry; 0.05 millihenry. (Note that the difference of these values divided by 4 gives the value of M found above.)

10. Calculate the energy in a coil whose self-induction is 180 henries when a current of 6 amperes flows through it. *Ans.* 3240 joules.

11. Show that because the self-induction of a coil having an air or other non-magnetic core varies as N^2 , it follows that the energy at constant impressed e.m.f. is independent of the number of circular turns of a given wire if all have the same circumference.

12. An ideal step-down transformer operates on a 2200-volt line supplying a load of 60 amperes. The winding ratio is 20 : 1. What are the primary current, secondary terminal volts, and power output? *Ans.* 3 amperes; 110 volts; 6.6 kilowatts.

CHAPTER 53

Electrical Machinery

723. The telephone. The receiver of the Bell telephone consists essentially of a permanent magnet NS , shown in Fig. 132, one end of which is wound with a coil C of many turns of fine wire, and a circular diaphragm D . This is supported around its periphery so as to lie very close to the pole without actually touching it. As a result, it is pulled toward the pole and thus is under a constant slight strain

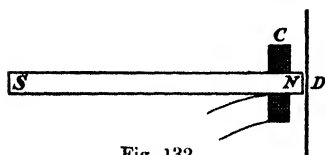


Fig. 132.

inward. Suppose the flux density B which passes through the diaphragm corresponds to the point p on the magnetization curve shown in Fig. 133, where the curve is steepest. Then a very slight change in H , represented by ΔH in the diagram, results in a relatively large change in B , shown as ΔB . Therefore if a very small current is sent through the coil C , a much greater effect is produced upon the diaphragm because of the flux B already passing through it than would be the case if the core were unmagnetized and the change in H caused by the current started at the origin. If the current through the coil increases and decreases periodically, the diaphragm is alternately attracted and released according to whether ΔH tends to increase or decrease the flux. These variations in the force attracting the diaphragm, if not too rapid, cause it to vibrate; and if the vibrations are of audible frequency, sound is emitted. The current required is of the order of 10^{-6} ampere, so the telephone receiver is a very sensitive detector of alternating currents.

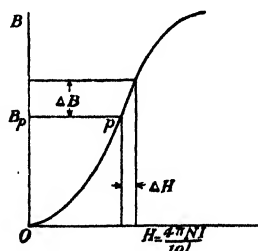


Fig. 133.

In the earliest types of telephones, an exactly similar device was used as the transmitter. The air vibrations set up by the voice impinge upon the diaphragm, cause it to vibrate, and thus vary the flux. These variations in the flux density induce a varying current in the coil that interlinks the flux. This current flows through the circuit connecting transmitter to receiver, and reproduces the original

vibrations in the receiving diaphragm, which acts upon the surrounding air to produce audible sound.

Such an arrangement needs no battery, but the e.m.f. induced in the transmitter is too small to set up currents of sufficient magnitude over the resistance of a long line. The Blake transmitter, and others using the same principle, make telephoning over long distances possible. Here the transmitter is really a **microphone**. This device is based on the remarkable fact that two carbon blocks touching each other offer widely varying resistance to currents passing between them, according to how closely they are pressed together. When such a contact forms part of an electric circuit containing a telephone receiver and battery, a slight jar of the contact is enough to make quite a loud sound.

This is usually demonstrated as shown in Fig. 134, where a carbon pencil is loosely supported between two grooved carbon blocks, so that the current must pass through two variable contacts in series.

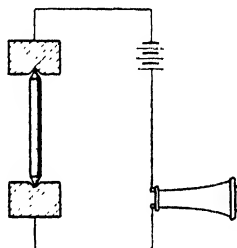


Fig. 134.

In the Blake transmitter this arrangement is made even more sensitive by using carbon granules packed together in a chamber upon which the diaphragm acts, compressing them more or less as it vibrates. The battery in the circuit sends varying currents through the varying resistance offered by the granules, and these currents flow through the primary of a small open-core transformer. The secondary has more turns than the primary, and a relatively large e.m.f. is induced in it. This e.m.f. is alternating because, when the primary current is increasing, it induces a secondary e.m.f. in one direction, and when it is decreasing, the induced e.m.f. is in the opposite direction. The alternating currents set up by the secondary e.m.f. operate the receiver at the other end of the line as has already

been explained.

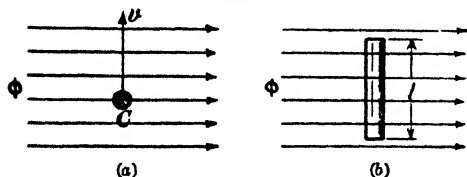


Fig. 135.

724. The generator. A dynamo-electric machine that develops electrical power at the expense of mechanical energy supplied to it is called a **gen-**

erator The fundamental principle by which this is effected has already been explained in Article 713. But certain modifications of the elementary equations may be made to adapt it to the par-

ticular case we are considering. Suppose a conductor C of length l moves with a velocity v across a flux ϕ , as shown in section and plan in Fig. 135, (a) and (b). Then the e.m.f. developed between the two ends of the conductor is given by $E = -d\phi/dt = -B(dA/dt)$, where dA is the area swept out by the moving inductor in the time dt . Let ds be the distance the inductor moves normal to the flux in the time dt . Then $dA = lds$, and
$$E = -Bl \frac{ds}{dt} = -Blv,$$

because ds/dt is its instantaneous velocity v . If instead of moving perpendicularly to the field, the inductor cuts it at some other angle α , as in Fig. 136, the rate of cutting the lines of force is reduced, and

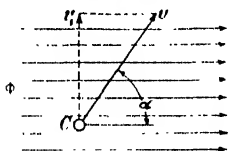


Fig. 136.

we must now consider the vertical component v_1 of the velocity, which alone is effective. Therefore a more general expression is

$$\begin{aligned} E &= -Blv \sin \alpha \text{ e.m.u.} \\ &= -\frac{Blv}{10^8} \sin \alpha \text{ volts.} \end{aligned}$$

If the velocity is uniform, this expression is valid at any instant. If not, it refers to the instantaneous e.m.f., or if v is the average speed, it gives us the average value of E .

725. Calculation of generator voltage. Suppose an iron drum is rotated clockwise around its axis O between the poles of a magnet N and S , as shown in Fig. 137, and let there be N inductors on its periphery, whose sections are shown as small circles. These inductors are connected in series so as to form a closed circuit by connections not shown, and which do not cut the flux. They may therefore be neglected as far as the calculation of the induced voltage is concerned.

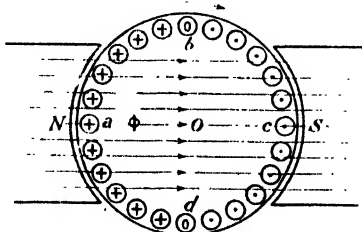


Fig. 137.

Each inductor is the source of some e.m.f., but as they cut the flux at different angles, varying from 90° at a and c , to 0° at b and d , their instantaneous voltages, as given by the preceding equation, are different. However, the *average* e.m.f. of each in passing from d to b , or from b to d , is the same for all. When an inductor goes from d to b , it cuts all the flux once in half the time T of one revolution. Then the average e.m.f. developed is

$$E_{av} = \frac{2\phi}{T} \text{ e.m.u.} \quad (1)$$

The negative sign previously used in calculating an induced e.m.f. is here omitted as it has no real significance in generator and motor problems.

In Fig. 137, it is clear that half of the total number N of inductors are developing an e.m.f. directed away from the observer, as indicated by the plus signs. The other half are developing an e.m.f. directed toward the observer, as indicated by the dots. Thus the total e.m.f. developed in each half of the armature is given by

$$\begin{aligned} E &= \frac{2\phi}{T} \times \frac{N}{2} \\ &= \frac{\phi N}{T} \text{ e.m.u.} \end{aligned} \quad (2)$$

This relation may be stated in a more useful form by introducing the speed n in revolutions per second or per minute. If it stands, as is quite usual, for r.p.m., then dividing (2) by 10^8 , and substituting $60/n$ for T , we obtain

$$E = \frac{\phi n N}{60 \times 10^8} \text{ volts.} \quad (3)$$

We may think of the armature as having two equal and opposite

voltages developed in its two halves, like two similar batteries connected in opposition, as in Fig. 138 (b). In such an arrangement no current flows around the closed internal circuit. But if we connect the points b and d through an external circuit closed by the key K , a

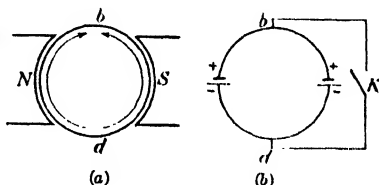


Fig. 138.

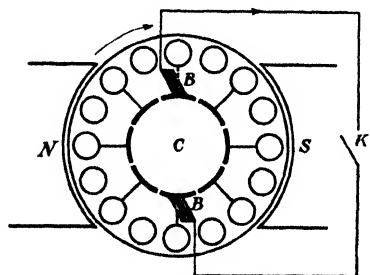


Fig. 139.

current will flow, each battery delivering half. The same could be done for the generator by placing "brushes" at b and d , thus forming rubbing contacts with the inductors as they pass under the brush. But this method would involve too much friction and a gradual wearing away of the winding, so instead, the inductors are tapped out at regular intervals to bars on a "commutator," shown at C in Fig. 139. The commutator is rigidly fixed to the armature shaft and rotates at the same angular velocity. Having a much smaller diameter, the "bars" move at a lower speed than

the inductors, and are designed to withstand wear, and present a smooth surface to the brushes BB (usually made of carbon) which bear upon them. Whether the upper brush is positive or negative depends upon the nature of the winding, the direction of the flux, and the sense of rotation. If it is positive, the current flows through the external circuit when the switch K is closed, as indicated by the arrow, and each inductor carries half of the total external current.

726. Armature reactions. The resistance R_a of the armature winding, as measured between brushes, is the resistance of its two halves in parallel. When the armature is delivering current, this resistance causes an internal drop of potential which lowers the terminal e.m.f. This drop is easily calculated when I and R_a are known, for the resistance of each half is $2R_a$, and the current is $I/2$; hence the internal fall of potential is IR_a , and the terminal e.m.f. between brushes is given by $E_t = E - IR_a$, where E is the entire e.m.f. generated.

The total electrical power developed by the armature is EI , and it would take an exactly equal amount of mechanical power to drive it if there were no friction or similar losses. Therefore, on open circuit, with no current in the armature inductors, it would take no power except that needed to excite the field and to overcome friction and certain other small fixed losses.

It requires power to run a loaded generator, because as soon as a current flows through its inductors, they exert a drag which opposes the motion, like the drag due to eddy currents. This might be called a motor reaction to the generator action, and it is demanded by the law of the conservation of energy. This reaction increases with the amount of current drawn from the generator, and the engine that drives it has to supply increasing power to overcome the increasing drag on the inductors.

In addition to friction, generator losses are due to wind resistance (windage), hysteresis and eddy-current losses in the armature core, and resistance losses. The total input is equal to the useful output $E_t I$ plus the losses L , and the efficiency is $E_t I / (E_t I + L)$. The most important items in L are the armature and field $I^2 R$ losses, which appear as heat. In order to compute them we must know their respective currents, the calculation of which is explained below.

727. The shunt generator. The field of a generator is usually produced by an electromagnet excited by current from the armature. This may be done in two ways: In **shunt generators**, the coil which produces the magnetism consists of many turns of rather fine wire connected across the brushes, and at all times when the generator is

running, a field current I_f , which is much less than the normal full-load current, is drawn from the armature. Therefore $I_a = I + I_f$, where I_a is the total armature current and I is the external current.

We may illustrate the current values in a shunt generator by the circuits shown in Fig. 140. Let the full load terminal e.m.f. be 110

volts. Let the field resistance be 220 ohms and let the external resistance be 10 ohms. Then the field current is 0.5 ampere, the main current is 11 amperes, and the armature current is 11.5 amperes. If the armature resistance is half an ohm, the armature drop is $11.5 \times 0.5 = 5.75$ volts;

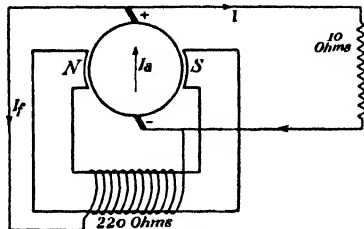


Fig. 140.

hence a total of $110 + 5.75 = 115.75$ volts is actually being generated by the armature. On open circuit, the armature drop is $I_f R_a$ because only the field current flows through it. This potential drop is $0.5 \times 0.5 = 0.25$ volt, if we assume the field current to be the same as before. Then the terminal e.m.f. with no external load on the generator would be $115.75 - 0.25 = 115.5$ volts. This is not quite the case however, because the field current is greater than before, owing to the increased voltage. This means an increased flux and consequently a still greater e.m.f., coupled with a slightly increased internal drop. The calculation is too complicated to discuss further here, but at least it is evident that the terminal voltage of a shunt motor varies with the load current, and that it must decrease with increasing load.

728. The series and compound generators. Another way of connecting the field of a generator is in series with the armature and external circuit.

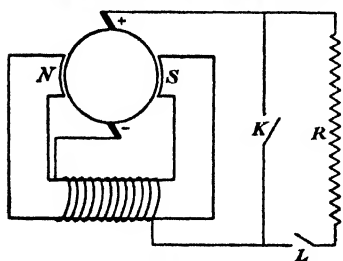


Fig. 141.

Then the load current flows through a few turns of heavy wire wound about the field core, as indicated in Fig. 141, and the generator is said to be **series wound**. At no load (open circuit) there is no field current, and the terminal e.m.f. is nearly zero, but not quite, because a feeble residual magnetism persists and enables the armature to develop

a slight e.m.f., as is also true of the shunt generator, before its field has "built up." If now the short-circuiting switch K is closed, this feeble voltage sends a current through the field coils, which thus

develop an added flux, and this in turn builds up E . So the machine rapidly develops the full-load current for which it is designed. Then the load switch L may be closed and K opened, when the generator delivers current to the load R . This building-up process occurs also when a shunt generator is started, but in this case its voltage must be built up on *open* circuit.

The two methods of winding just described are often combined in **compound-wound** generators, which have both shunt and series field coils. The latter is designed to keep the terminal e.m.f. nearly constant at all loads by supplying an added flux when the shunt field would otherwise tend to fall off with increasing armature potential drop, as has been explained. The series turns carry an increasing current with increasing load, which offsets the tendency of the shunt field toward decreasing flux.

729. Alternating current generators. If a coil of N turns of wire, shown in section as CC in Fig. 142, is rotated around an axis X which

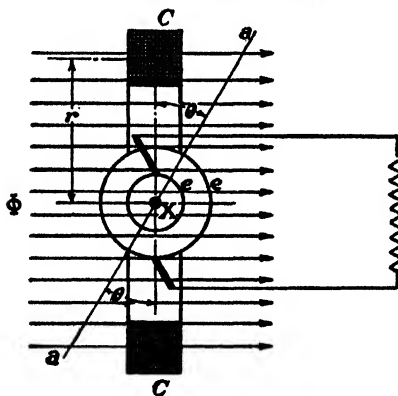


Fig. 142.

is perpendicular to a uniform flux, an alternating e.m.f. will be developed in its turns. This may be made available by connecting the two ends to collecting rings ee , which rotate with it like the commutator already described, and on which bear two brushes leading to the external circuit. As the coil rotates, it cuts the flux at a rate which varies as the sine of the angle it makes with the vertical. Thus when the coil is at aa , it makes

an angle θ with the vertical CC . When $\theta = 0$, the coil is vertical and embraces the maximum amount of flux, but its rate of cutting the flux is zero. At right angles to CC , the coil embraces no flux, but as $\sin \theta$ is unity, the rate of cutting the flux is a maximum. As the induced e.m.f. varies as the rate at which the flux is cut, it too varies as the sine of the angle θ .

During one complete revolution, the coil cuts the entire flux four times. As θ increases from zero to ninety degrees, the interlinking flux falls from maximum to zero. From 90° to 180° it increases to a maximum again. In the third quadrant, the interlinking flux falls once more to zero, but recovers the maximum value at the end of a

complete revolution. Therefore, in the time T of a revolution, ϕ lines of force have been cut four times by N turns, and the average e.m.f. induced is

$$E_{av} = \frac{4\phi N}{T} \text{ e.m.u.} \quad (1)$$

As explained above, the e.m.f. varies as the sine of θ . Then its instantaneous value is given by

$$e = E_m \sin \theta, \quad (2)$$

where E_m is the maximum value or amplitude of e . In each cycle, e reaches E_m twice: when $\theta = \pi/2$, and when $\theta = 3\pi/2$ radians. Now by a well-known theorem of trigonometry, the average value of the sine of an angle is $2/\pi$. Therefore $E_{av} = 2E_m/\pi$, or

$$E_m = \frac{\pi}{2} E_{av}. \quad (3)$$

Hence, taking E_{av} from (1), we obtain

$$E_m = \frac{2\pi\phi N}{T} \text{ e.m.u.}, \quad (4)$$

and substituting this value in (2), we have

$$e = \frac{2\pi\phi N}{T} \sin \theta \text{ e.m.u.},$$

or

$$e = \frac{2\pi\phi N n \sin \theta}{60 \times 10^8} \text{ volts}, \quad (5)$$

where n represents revolutions per minute. In the preceding equations we may always calculate the total flux ϕ which interlinks the coil, provided we assume its thickness to be negligible in comparison with its radius r . Then $\phi = \pi r^2 B$, where B is the flux density due to the field magnets.

The alternating e.m.f. developed by the rotating coil may be plotted as a function either of θ or of the time, because $\theta = \omega t$, where ω , as usual, is the angular velocity $2\pi n$. This graph is shown in Fig. 143, with the X axis laid off in fractions of the period and also in radians.

The rotating ring just described is essentially an alternating-

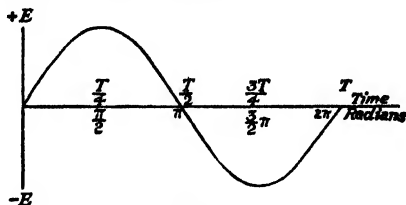


Fig. 143.

current generator. In practice the field is produced by an electromagnet which is excited from a source of direct currents such as a storage battery or small direct-current generator, called an "exciter." As the current needed by the field coils is much smaller than the line current supplied by the armature, it is customary to have the field rotate within a stationary armature, the two parts being called **rotor** and **stator**, respectively. The field current is fed to the field coils through brushes bearing on collecting rings which are mounted on the same shaft as the rotor and revolve with it. The main current is then drawn from the stator without rubbing contacts.

730. Alternating currents. The current developed by an alternating e.m.f. alternates also, though it is not necessarily of the same wave form. In what follows, however, we shall assume that both E and I vary sinusoidally. In measuring these quantities we have seen that we may concern ourselves with either their instantaneous, maximum, or average values. But there is a fourth aspect, known as the *effective* value, which is even more important. In the case of the current, it is the same value as that of a direct current which would produce the same I^2R heat in a resistance as is actually produced. In the case of the e.m.f., it is the same value as that of a direct e.m.f. which, when multiplied by the effective value of the current, gives the true power EI consumed in the circuit that has no counterelectromotive force. This can be shown to be the square root of the mean square of the variable I or E , and when these quantities vary sinusoidally, it is found to be their maximum values divided by $\sqrt{2}$. Thus $I_e = I_m/\sqrt{2}$, and $E_e = E_m/\sqrt{2}$. The subscript e , denoting *effective*, is usually omitted, as the effective values are always understood, unless otherwise stated. Taking the maximum value of the induced e.m.f. from equation (4) in Article 729, we find that the effective value in the case of this particular machine is given by

$$\begin{aligned} E &= \frac{\sqrt{2}\pi\phi N}{T} \text{ e.s.u.} \\ &= \frac{\sqrt{2}\pi\phi Nn}{60 \times 10^9} \text{ volts} \\ &= \frac{7.4\phi Nn}{10^{10}} \text{ volts.} \end{aligned} \tag{1}$$

Unless an A.C. circuit is composed wholly of noninductive resistance, the current does not in general pass through its maximum

and zero values at the same time as its e.m.f. This involves a lag or a lead in phase. A current lagging one eighth of a period, or $\pi/4$ radians, is shown in Fig. 144 (b). This may be conveniently represented by the two vectors E and I , as in (a). These are supposed to develop the two sine curves by a counterclockwise rotation, their lengths being equal to the maxima of the two curves. In measuring

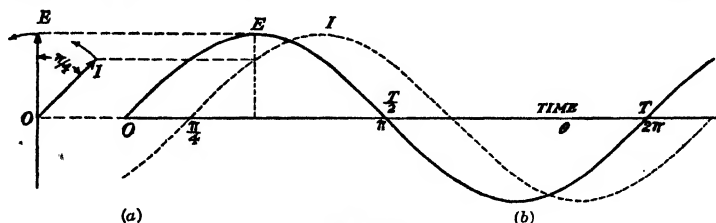


Fig. 144.

power, we must take account of the fact that I is sometimes on the same side of the time axis as E , and sometimes opposite. When both E and I are positive, or both negative, the instantaneous value ei of the power is obviously positive, and is being supplied to the circuit, but when they are of opposite signs, the product ei is negative, and power is being returned by the circuit to the generator or other source. An analysis of this state of things shows that the actual power delivered is given by $EI \cos \phi$, where ϕ is the phase angle between E and I . The cosine of ϕ is called the power factor, and it varies between unity (no lag) and zero (I lagging or leading E by 90°). In general, self-induction in a circuit causes the current to lag, as in Fig. 144, while capacitance causes it to lead. When both are present they may neutralize each other and produce a power factor of unity, as is also the case with a noninductive circuit containing only resistance. To calculate the value of the alternating current in a given circuit, and its angle of lag, calls for a more advanced treatment of this somewhat complicated problem. But it may be noted that when the only work done appears in the form of heat, as in lighting circuits, the power is always equal to I^2R ; therefore

$$EI \cos \phi = I^2R. \quad (2)$$

If, in addition, there is no inductance or capacitance in the circuit, $\phi = 0$, $\cos \phi = 1$, and $EI = I^2R$. In this special case (but not otherwise) Ohm's law applies, as is evident when we divide the equation by I . Thus with a noninductive load of resistance units, such as lamps, alternating currents may be calculated as if they were direct.

Motors

731. General equation. It was demonstrated in Article 724 that when a conductor moves across a uniform magnetic flux of density B with a velocity v , an e.m.f. equal to $-Blv$ is set up between its ends. But in this discussion, as in Article 725, we shall ignore the negative

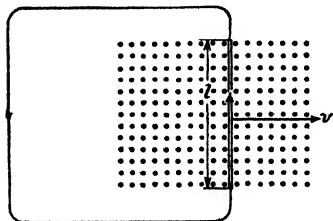


Fig. 145.

sign. Then if the circuit is completed by a wire which lies entirely outside the field, a current Blv/R flows through a conductor whose length is l and whose velocity is v , as shown in Fig. 145. This current calls for a certain amount of work which must be done on the conductor in moving it against the force F that

opposes the motion. Therefore $W = Fs$, where s is the distance moved at right angles to both wire and flux. Then, equating the electrical energy expressed in absolute units (ergs) to the mechanical work done, we have $Fs = EIt$, and substituting Blv for E , we obtain

$$Fs = Blvit.$$

But

$$v = s/t.$$

$$\therefore F = BI l,$$

or

$$F = BI l / 10 \text{ dynes,} \quad (1)$$

when I is measured in amperes. Since $B = H$ in air, (1) may also be written $F = HI l / 10$, as was proved by another method in Article 621. This relation gives the sidewise push on the conductor, and is the fundamental equation of the electric motor.

732. The motor and torque. The wiring of the direct-current motor is essentially the same as that of the generator. Reduced to its basic principles, the motor armature is an iron ring carrying inductors on its periphery. These are joined in series by "end connections," which do not cut the flux

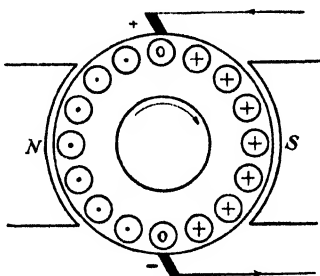
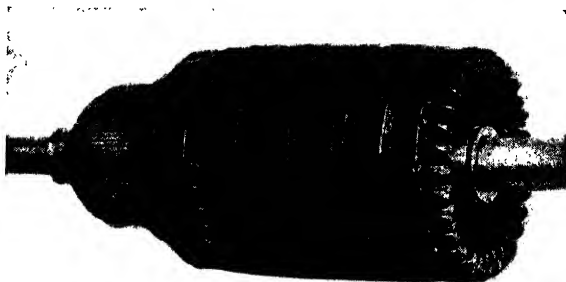


Fig. 146.

as the armature rotates. The current enters the armature by means of brushes bearing on a commutator, though in Fig. 146, for the sake of simplicity, they are shown bearing on the inductors. If the upper brush is connected to the positive terminal of the source of current,

and if the armature is wound in the same "sense" as that of the generator shown in Fig. 137, then the current which divides at the brush goes through each half of the winding in the opposite direction from that of the generator currents, as is indicated by the plus signs and dots. This is because the current is now supplied from an external source and flows from the positive to the negative brush through the



Courtesy General Electric Co.

Plate 18.

Armature and commutator of a D.C. motor, showing "end connections" for drum winding.

armature, while in the generator it flows internally from negative to positive, as in all sources of electromotive force.

As in a generator, the armature rotates between the poles of an electromagnet. If the flux is horizontal, these act with a vertical force equal to Bil on each inductor. The direction of this force may be determined by the fingers of the *left* hand used with the same significance as those of the right hand in the case of the generator. As a result of these forces, a torque is developed which we find is *in the same sense* as that which was applied to the generator, causing it to rotate against the opposing drag of its inductors. This may be illustrated as follows: Suppose the shunt generator shown in Fig. 147 delivers current to a line whose voltage is maintained by other generators in

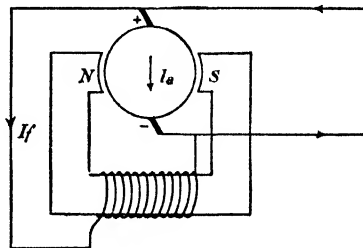
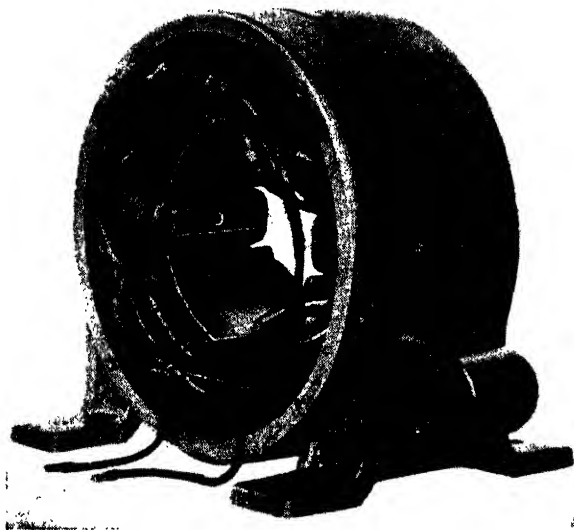


Fig. 147.

parallel with it. Then if its e.m.f. falls below that of the line, the current through its armature reverses. But since the field is shunted across the line, it does not reverse, and the dynamo begins to operate as a motor without changing the sense of rotation.

If it is desired to reverse the sense of rotation of a motor, the connections of the field are interchanged at the brushes. Then the flux reverses, but the armature current does not, and the motor runs backward. This argument also applies to a series-wound motor, because the field and armature currents reverse simultaneously when the main terminals are interchanged, and the direction of rotation remains



Courtesy General Electric Co.

Plate 19.

Wound field frame for four-pole D.C. motor, using armature shown above. The narrow "inter-poles" between the four large ones are to prevent sparking at the brushes under varying load.

unchanged. But reversing the field terminals alone results in a reversal of the sense of rotation.

733. Operation of the shunt motor. As soon as a motor begins to rotate under the action of the current supplied to it, the inductors begin cutting the field flux, and an e.m.f. is set up in them, which, according to Lenz's law, must tend to oppose the motion that caused it. This generator action of the motor is a corollary of the motor action of the generator, already mentioned in Article 726, and it appears as a back e.m.f. tending to oppose the current. Therefore Ohm's law, applied to the armature current, must be modified to read

$$I = \frac{E - E'}{R_a}, \quad (1)$$

where E' , the back e.m.f., is zero when the motor is at rest. It increases steadily as the motor speeds up, approaching but never equaling E when the motor approaches full speed. This fact shows that the current tends to decrease with increasing armature speed, from a very large initial value when $E' = 0$, up to a very small one when the motor runs at full speed with zero load. In order to protect the armature from excessively large and dangerous initial currents, a starting rheostat in series with the armature is used, as shown in Fig. 148.

Suppose an armature A has a resistance of 0.5 ohm, and that it develops a back e.m.f. of 105 volts when running at normal speed on a 110-volt circuit. Then, without a starting rheostat, a current of $110/0.5 = 220$ amperes would flow with the armature at a standstill, but only $(110 - 105)/0.5 = 10$ amperes flow at full speed. Therefore a

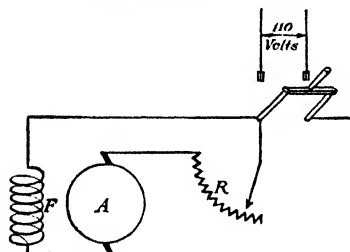


Fig. 148.

rheostat of say 5 ohms should be used, which cuts down the initial current to twice its running value, as seen from $I = 110/(5 + 0.5) = 20$ amperes.

734. Efficiency and torque. The equation $I = (E - E')/R_a$ may be transposed to read $IR_a = (E - E')$, and if both sides are multiplied by It , we obtain

$$I^2 R_a t = (E - E') It, \quad (1)$$

which is the armature heat loss expressed in joules. This may be written $EIt = H + E'It$, where EIt is the energy supplied to the armature during the time t , and H is the $I^2 R_a t$ heat loss. The quantity $E'It$ is equal to the difference between these two quantities, and must be the energy output of the motor, provided all other losses but H are neglected. As H is much the largest loss, this is permissible in a rough estimate. Therefore the energy output of a motor is $W = E'It$ joules approximately, and the power P developed is $E'I$ watts.

Since the input is at the rate of EI watts, the efficiency, which is the ratio of output to input, becomes $E'I/EI = E'/E$. The motor described above would have an efficiency of $(105/110) \times 100 = 95.5$ per cent. This simple calculation has been based on the assumption that losses other than the armature heat loss could be ignored. However, some of them are really not negligible, such as the $I^2 R$ loss in the field coils, the losses due to bearing and wind friction, and the

core losses due to hysteresis and eddy currents. These total a value which may lower the ideal efficiency by several per cent.

The torque developed by the ideal motor is calculated as follows: The mechanical power of any rotor is given by the product of the torque L and the angular velocity ω , or $2\pi n/60$, where n is the number of revolutions per minute. Therefore, equating this product to the useful electrical power $E'I$, we have $E'I = 2\pi nL/60$, or $L = 60E'I/2\pi n$, showing that the torque at a given speed depends upon the armature current and the back e.m.f. As the latter is identical with the generator e.m.f. (Article 725), we may substitute $\phi nN/(60 \times 10^8)$ for E' , and reducing to absolute units obtain

$$L = \frac{\phi NI}{2\pi 10^8} 10^7 \text{ dyne-cm} \quad (2)$$

which is thus seen to be independent of n , as would be expected, and to vary only with the product ϕI for an armature of a given winding.

735. Motor regulation. Since $E' = \phi nN/(60 \times 10^8)$, we may substitute this value in equation (1), Article 733. Then the current taken by a motor may be expressed by

$$I = \frac{E - kn}{R_a}, \quad (1)$$

where $k = \phi N/(60 \times 10^8)$. Now if the motor is running at full speed and no load, the back e.m.f., kn , has its maximum value, and I is a minimum, being just large enough to supply the required no-load torque. But as the motor is increasingly loaded, n decreases enough to admit the necessary additional current, very much as a valve is opened by the governor to admit more steam to a loaded engine.

As a concrete illustration of motor regulation, suppose the torque under full load to be five times as great as the no-load torque. Then the current must also increase five times. If the new speed is n' r.p.m., equation (1) becomes

$$5I = \frac{E - kn'}{R_a}. \quad (2)$$

Dividing (2) by (1), we obtain $E - kn' = 5(E - kn)$, whence

$$4E = k(5n - n'). \quad (3)$$

If the values of E and E' (at no load), as assumed in the last paragraph, are 110 and 105 volts respectively, and if the speed n is 1260 r.p.m., then $k = E'/n = 105/1260 = 1/12$, and equation (5) becomes

$$4 \times 110 = \frac{5 \times 1260 - n'}{12}, \text{ and } n' = 1020 \text{ r.p.m.,}$$

which is a 19 per cent fall in the speed corresponding to a 500 per cent increase in the power developed.

736. Changing the speed of a shunt motor. The constant k by which we calculate the back e.m.f. of a motor is equal to $\phi N / (60 \times 10^8)$. Therefore if the flux is increased by strengthening the field current, k increases in the same proportion. But since the back e.m.f. is not much smaller than the impressed, we may write the approximate relation $E = E' = kn$, or $n = E/k$. Therefore, strengthening the field flux by increasing k decreases the speed n . Conversely, weakening the field flux speeds up the motor. This is true, however, only within the limits where E and E' do not differ too much. If the load is abnormally great, and n is reduced in a very large proportion, then the above assumption is no longer valid. In such a case, weakening the field may reduce the flux so much that the motor will "stall" instead of speeding up, whereas strengthening the field under these conditions would accelerate it. The critical load which gives rise to this state of things may be calculated for a given motor by a somewhat complicated analysis.

737. Field of the induction motor.

The most common form of alternating-current motor is the induction motor, so called because the armature current is induced by the field that develops a rotating magnetic flux. In the two-phase induction motor the coils aa (Fig. 149) are in series with phase A , and are so connected that when the current in that phase is a maximum, the two halves of the ring are magnetized in opposition.

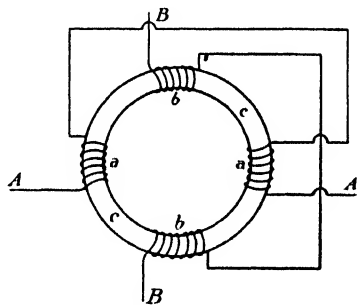


Fig. 149.

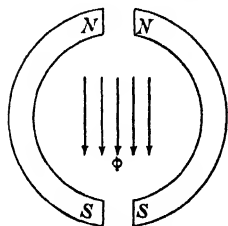


Fig. 150.

are formed at the top and bottom, as if two U-magnets were placed with like poles opposite to each other, as shown in Fig. 150. Phase B is also an alternating-current circuit, but in quadrature with phase A , which means that it leads A or lags behind it by 90° , or a quarter period. Thus when A is a maximum, B is zero, and vice versa. At the time t_1 , indicated in Fig. 151, there is no flux but ϕ_A , and the resultant poles are as shown above. But when t_2 is reached, both coils have equal currents flowing in them. Then the resultant poles are due

both to ϕ_B at aa , Fig. 149, and to ϕ_A at bb ; thus an effective flux is produced between the points cc . At t_3 , ϕ_A is zero, the resultant poles

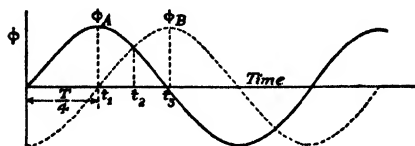


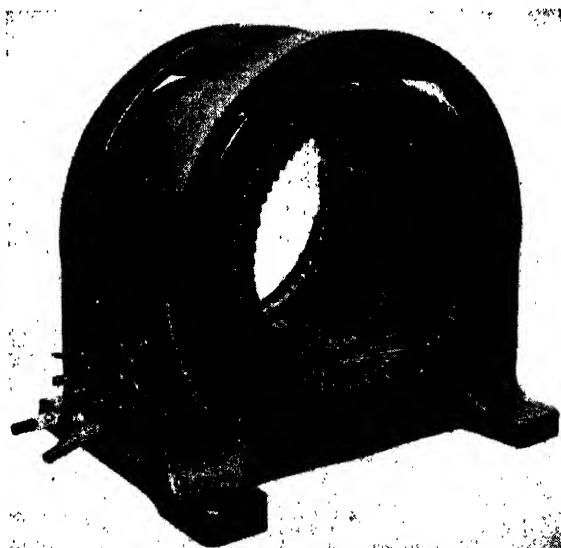
Fig. 151.

are due to ϕ_B alone, and are at aa . Since the total flux is the vector sum of two fluxes at right angles to each other, and since one lags a quarter period behind the other in time, the resultant flux at any instant

may be calculated from $\phi_A = \phi_m \sin \theta$, and $\phi_B = \phi_m \sin (\theta - 90^\circ)$, where ϕ_m is the maximum value of the flux. Then the instantaneous resultant flux is given by

$$\phi_R = \sqrt{\phi_A^2 + \phi_B^2} = \phi_m \sqrt{\sin^2 \omega t + \cos^2 \omega t} = \phi_m.$$

Therefore the resultant flux is constant in magnitude and equal to



Courtesy General Electric Co.

Plate 20.

Stator (field) of a 3-phase, 60-cycle, 220- or 440-volt induction motor. The field rotates at 1200 r.p.m.

the maximum value of the flux due to either phase. This flux makes one complete revolution around the field ring for a cycle of the alternating current.

738. Armature of the induction motor. Actually any metallic disc would be set spinning by the rotating field described in the last article, because the eddy currents induced in it would produce a drag tending to make it keep up with the moving flux, as explained in Article 721. But this would be very inefficient. Therefore armatures are designed in which the induced currents follow well-defined paths, as in direct-current motors. The most usual type is the "squirrel cage" armature. The core is of laminated iron, which makes a good path for the rotating flux, but does not permit eddy currents of any magnitude. Copper bars are imbedded in slots around the periphery, and these are short-circuited at their ends by two copper rings. This arrangement, without the iron core, is illustrated in Fig. 152.

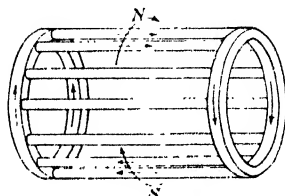
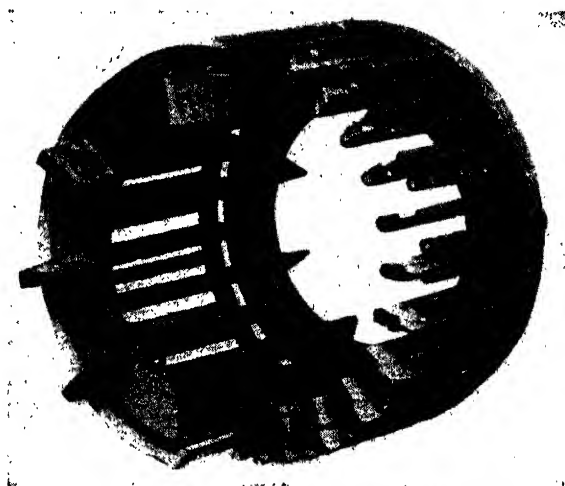


Fig. 152.



Courtesy General Electric Co.

Plate 21.

Sectional view of "winding" of squirrel-cage rotor formed by casting around laminated core. In the photograph, the core has been removed by acid.

As the flux sweeps across the inductors, it induces currents which flow by way of an end ring to other inductors in which an e.m.f. is being developed that favors the current at the moment. Then the circuit is completed by the ring at the other end. This distribution of

currents is indicated by the arrows in the diagram. The curved outer arrows show the direction in which the resultant field poles must be moving if the directions of the induced currents are to be as indicated.

If there were no load on the armature, and if it revolved without friction or other losses, it would rotate at the same speed as the flux. In this case no e.m.f. would be induced and no currents would flow along the conducting bars. But if a load is put on, the armature slows down just enough to allow sufficient relative motion between bar and flux to induce just the necessary amount of torque required by the load. On account of friction, hysteresis, and other losses, the no-load speed is not quite synchronous with the field, and when loaded, the armature rotates at a still lower rate. But as in the case of the shunt motor, the relative retardation or "slip" is small compared to the proportional increase in the load, and for small variations of the load, the speed is nearly constant.

739. Forces between currents. Two parallel wires carrying current in the same direction are pulled together by the mutual action of their magnetic fields. This fact was discovered by Ampère in his classic experiments in 1820, only a few months after Oersted had announced his great discovery. The reason for this force becomes evident from an examination of Fig. 153. The two sets of concentric circles representing the fields interlink the two wires shown in section at *A* and *B*. These wires carry parallel currents flowing directly

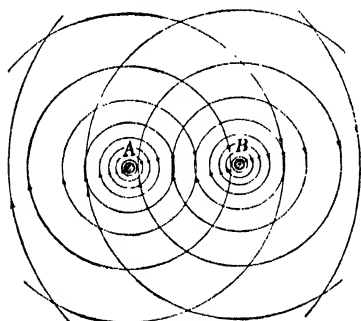


Fig. 153.

away from the observer. The two fields are evidently in opposition between the wires, but reinforce each other outside. This results in a field between the wires which is weaker than that which surrounds them both. But as lines of force behave like stretched elastic cords, the result is to force the wires together. If the currents were directed oppositely to each other, there would be an intensification of the field between them and a weakening outside, resulting in repulsion.

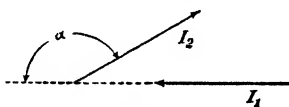


Fig. 154.

If the conductors are inclined at any angle α to each other, the attracting force varies with $\cos \alpha$, as is evident from Fig. 154, where α is greater than 90° , its cosine is negative, and repulsion takes place.

If α is 90° , the cosine is zero, and there is no attraction or repulsion. If the angle is less than 90° , the cosine is positive and the conductors attract each other.

In addition to attraction and repulsion, there exists in general a torque between two crossed wires, which tends to make them parallel with the currents flowing in the same direction. It is zero when α is zero, reaches a maximum at 90° , and falls to zero again when α is 180° .

SUPPLEMENTARY READING

N. E. Gilbert, *Electricity and Magnetism* (Chap. 16), Macmillan, 1932.

A. Zeleny, *Elements of Electricity* (Chap. 18), McGraw-Hill, 1930.

W. H. Timbie, *Elements of Electricity* (Chapters 7, 8, 9), Wiley, 1925.

PROBLEMS

1. An inductor bar of a generator is 80 cm long, and is 120 cm from the axis of the armature which drives it. The flux density which it cuts is 12,000 gauss, and the armature makes 1200 r.p.m. What is the induced e.m.f. when the angle α (Fig. 136) is 60° ? *Ans.* 125.3 volts.

2. Calculate the e.m.f. of a two-pole generator whose armature has 400 inductors cutting a flux of 8×10^5 maxwells per pole at 2400 r.p.m. *Ans.* 128 volts.

3. The armature resistance of a shunt-wound two-pole generator is 0.3 ohm. It develops 120 volts on open circuit. What is its terminal e.m.f. when a current of 25 amperes flows through the armature? What is its efficiency, considering only the armature loss? *Ans.* 112.5 volts; 93.8 per cent.

4. In the generator of Problem 3, the field has a resistance of 240 ohms. What is the I^2R loss when it is running on open circuit? *Ans.* 60.08 watts.

*5. Assuming that the e.m.f. actually generated by the dynamo of Problems 3 and 4 remains constant with light loads, calculate the field and armature currents, the terminal volts, the total heat loss, and the efficiency, if the external current is 10 amperes, and the fixed losses are 50 watts. *Ans.* 0.487 ampere; 10.487 amperes; 116.85 volts; 89.91 watts; 89.3 per cent.

6. Calculate the maximum and the effective values of the e.m.f. developed by an A. C. generator, constructed as shown in Fig. 142, when the flux is 5×10^5 maxwells, the angular velocity is 1800 r.p.m., and when there are 300 turns of wire. *Ans.* 282.6 volts; 200 volts.

7. The generator in Problem 6 delivers 18 amperes to a load in which it lags 30° behind the impressed e.m.f. What is the actual power? What is the load resistance if the circuit contains only coils of wire? *Ans.* 3118 watts; 9.62 ohms.

8. What is the force in kg which acts upon a metal bar 60 cm long in a transverse field whose intensity is 12,000 gauss, when a current of 20 amperes is flowing in it? *Ans.* 1.47 kg.

9. A rectangular coil of 40 turns, whose section may be represented by Fig. 142, has an average length ($2r$) of 18 cm, and an average width of 15 cm. The flux density is 80 gauss. Calculate the force on each horizontal side of the rectangle when the current flowing in it is 4 amperes. Calculate the torque when $\theta = 90^\circ$, 30° , and 0° . *Ans.* 19,200 dynes; 345,600 dyne-cm; 172,800 dyne-cm; zero.

10. A bipolar shunt dynamo has an armature with 400 inductors which cut a total flux of 1 million maxwells per pole. When run as a motor at 1800 r.p.m., what is the back electromotive force that it develops? *Ans.* 120 volts.

11. The resistance of the armature in Problem 10 is 0.4 ohm, and the field resistance is 180 ohms. The impressed e.m.f. is 130 volts, and the speed is 1800 r.p.m. at full load. Calculate the full load armature and field currents, the total power input, the I^2R loss, and the efficiency both when the field loss is considered and when it is not. *Ans.* 25 amperes; 0.72 ampere; 3344 watts; 343.9 watts; 89.4 per cent; 92.3 per cent.

12. What should be the resistance of a starting rheostat if the motor of Problem 11 is to start without drawing more than full-load current? *Ans.* 4.8 ohms.

13. Calculate the torque of the motor of Problems 10 and 11 when it is running under the specified conditions. *Ans.* 15.92×10^7 dyne-cm.

*14. If the load on the motor of Problem 11 is decreased so that it speeds up to 1900 r.p.m., what are the back e.m.f., the armature current, torque, and approximate efficiency? *Ans.* 126.67 volts; 8.4 amperes; 5.35×10^7 dyne-cm; 97.4 per cent.

15. A shunt motor takes 10 amperes at no load on a 120-volt circuit. Its armature resistance is 2 ohms. What current does it take when loaded to run at 75 per cent of no-load speed? *Ans.* 22.5 amperes.

16. Calculate the speed at which the motor of Problem 11 must run at no load, if there are no fixed losses. *Ans.* 1950 r.p.m.

*17. If the no-load torque which overcomes the fixed losses of the motor of Problem 11 is 3×10^7 dyne-cm, what are the no-load current, speed, and back e.m.f.? *Ans.* 4.71 amperes; 128.1 volts; 1921.5 r.p.m.

*18. If the motor of Problem 17 is loaded by a torque of 18×10^7 dyne-cm, what are its speed and the per cent change from no-load speed? *Ans.* 1780 r.p.m.; 7.4 per cent.

19. Calculate the field strength 3 cm from a very long straight wire that carries a current of 15 amperes. What is the force with which it attracts a parallel wire 40 cm long, 3 cm distant, and in which a current of 20 amperes is flowing? *Ans.* One oersted; 80 dynes.

CHAPTER 54

Electrical Oscillations

740. Damped vibrations. Let the condenser C in Fig. 155 be connected in series with an inductance L of low resistance and a key K . Let it then be charged to a potential difference of V volts with K open. When K is closed, the condenser discharges through the circuit, as shown by the portion of the curve, A to B , in Fig. 156. The resulting current rises (a to b) to a maximum value and then begins to decrease (b to c). During the time it is increasing, the inductance tends to oppose its growth by a counter-e.m.f. given by $E' = -Ldi/dt$, as explained in Article 717. Similarly it retards the current's decay by an e.m.f. $E'' = +Ldi/dt$. Thus, because of its electromagnetic inertia, the current tends to keep flowing even after the condenser is discharged. This results in charging the condenser again, but in the opposite sense (B to C), and the charging process continues until the potential of the new charge becomes equal and opposite to E'' .

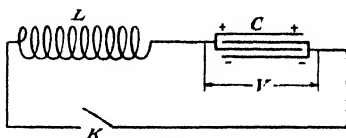


Fig. 155.

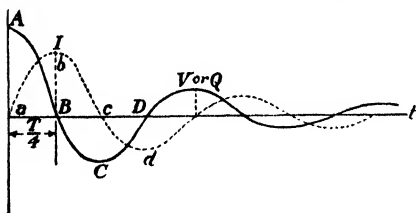


Fig. 156.

Then the current is zero and about to begin flowing in the opposite direction (c to d) as the condenser's reversed charge begins to flow back through the circuit.

On account of losses due to unavoidable resistance in the circuit, some energy is dissipated, and the potential obtained after reversal is less than at first. Therefore the oscillation is "damped," and dies down rapidly to zero after a series of steadily diminishing pulsations of the charge. This may be plotted with either V , Q , or I as a function of the time, giving the damped sine curves of Fig. 156. The solid line represents either V or Q , both of which are evidently in the same phase. The dotted line represents the current, which is zero when the charge on the condenser is a maximum, and therefore lags nearly 90°

behind the other curve. It would lag exactly 90° if there were no losses, but then the oscillation would be undamped, $\cos \phi$ would be zero, and the power consumed, $VI \cos \phi$, would be zero also.

This process has an almost exact mechanical analogue when a spring acts upon a rolling mass, as shown in Fig. 157. The elasticity

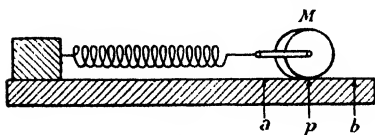


Fig. 157.

of the spring corresponds to the capacitance of the condenser, and the inertia of the mass corresponds to the inductance of the coil.

Suppose the cylindrical mass M rolls with very little friction upon the horizontal plane, and that the two ends of its axis are fastened to a spring whose other end is fixed to the block. If p is its position of rest with the spring neither extended nor compressed, and if it is stretched to b , the elastic tension on the spring represents the charge of the condenser. There is now no motion. This corresponds to the absence of current associated with an insulated charged condenser. Then if the roller is released, the spring contracts, and the velocity (current) increases until, when p is reached, the inertia of the mass keeps it moving, but with diminishing speed, to a point a . This motion, from p to a , compresses the spring until its elastic reaction overcomes the inertia of the moving mass. The velocity (current) is again zero. The compression of the spring corresponds to the reversed charge on the condenser, but because of friction losses, it is a little less than the original extension. Thus the roller oscillates back and forth with decreasing amplitude until it finally comes to rest.

The theory of damped oscillations gives the following value for the frequency:

$$n = \frac{1}{2\pi} \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}},$$

where R , L , and C are the resistance, self-induction, and capacitance of the oscillating circuit. If R is very small compared to L , the second term becomes practically zero, and the frequency may be approximately calculated from

$$n = \frac{1}{2\pi} \sqrt{\frac{1}{LC}},$$

or the period from

$$T = 2\pi \sqrt{LC}.$$

But if R is large and the resulting damping is great, the frequency is less than this simplified form of the equation indicates. If R is so large that $1/LC = R^2/4L^2$, or $R = 2\sqrt{L/C}$, then $n = 0$, and the

system does not oscillate. In this case the condenser simply discharges once, while the current sinks to zero. The resistance that is just large enough to produce this effect is called the **critical or damping resistance** of the circuit.

741. Electromagnetic waves. In Fig. 158, an oscillatory circuit is shown open at a spark gap G , composed of two brass knobs. The discharge is now accompanied by a spark across the gap, and it will oscillate in the same manner as the circuit as a whole. In fact, this is the simplest way of producing oscillations, for the condenser can be raised to the rather high potential required to break down the resistance of the gap before any discharge occurs. Experiment has shown that after G has once yielded to the electrostatic strain, there may be half a dozen or more oscillations before the potential of the condenser becomes too low to send a further discharge across the gap.

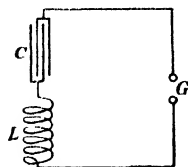


Fig. 158.

While the knobs of the gap are being charged, the lines of electrostatic force assume successive positions, as shown in Fig. 159 (a), and as the ends of the lines of force move along the rods, they are linked

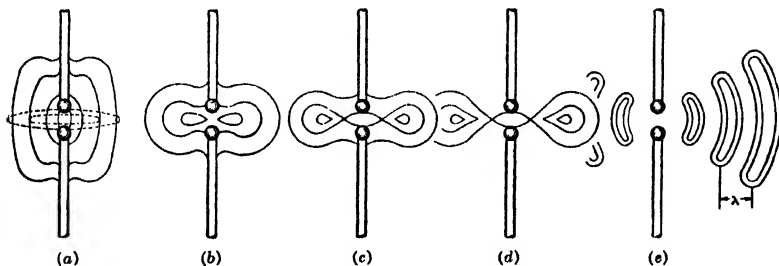


Fig. 159.

with a concentric magnetic flux, suggested by the horizontal dotted rings. Then the discharge begins, and the two ends of the lines meet to form a closed loop, as shown in (b), while (c) and (d) show other lines meeting to form other rings, and (e) shows the fully developed set of waves that has resulted from this oscillatory discharge.

The wave front at any moment is composed of electrostatic (Article 585) and magnetic fields (Article 559) at right angles to each other, and these determine a plane (normal to the paper) which is at right angles to the direction of propagation. Thus the three vectors, velocity, E , and H , are mutually perpendicular, as in the case of the corresponding quantities represented by the right-hand rule for the

induction of an e.m.f. As the spark oscillates, a continuous reversal of E occurs, and waves traveling with the speed of light in all directions, except along the axis of the gap, are thus sent out into space.

The diagram, Fig. 160, shows an electromagnetic vibration with H and E mutually perpendicular and in phase with each other. This

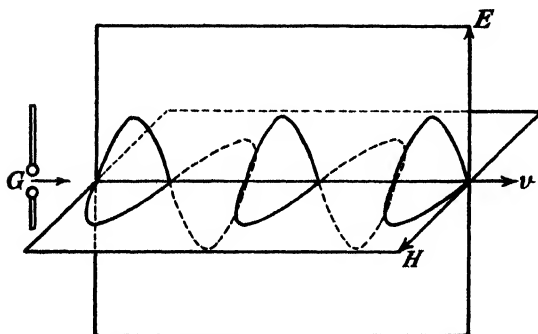


Fig. 160.

condition of phase is established a short distance away from the gap, although they are in phase quadrature at the start. The wave motion just described was deduced on theoretical grounds by Maxwell in 1864.

In a remarkably brilliant analysis,

he showed that light waves could be of this character, and predicted the exact nature of the much longer electromagnetic waves now used in radio communication.

During the years 1886 and 1887, Hertz, a German physicist, produced electromagnetic waves by discharging a condenser in the manner described above, and observed their presence at a distance. The detector consisted of a metal ring (Fig. 161) whose plane was normal to the direction of propagation of the waves, and which was equipped with a micrometer spark gap parallel to the original spark. When the discharge occurred, the induced currents were set up in the ring, resulting in another spark across its gap g provided its natural period of vibration, $T = 2\pi\sqrt{LC}$, was the same as that of the original discharge. The resonance thus established depended mainly upon the ring's radius. This determined the value of L , which had to be relatively large on account of the ring's very small capacitance.

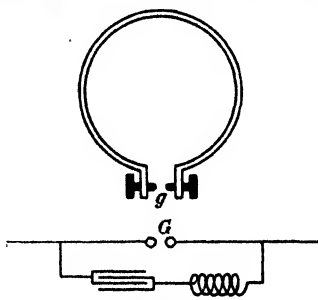


Fig. 161.

Hertz found that the waves predicted by Maxwell behaved in every respect like light, though those he investigated were very much longer,

being of the order of half a meter or more in length. They could be reflected, refracted, diffracted, and polarized with suitable apparatus, so that Maxwell's theory received complete vindication and has since been regarded as the best explanation of all so-called ether waves.

742. Wireless telegraphy. Hertz's experiments, though of vital theoretical interest, involved such small amounts of energy and such insensitive detectors that they did not immediately suggest the possibility of signaling to a great distance without wires. But Marconi, about ten years after Hertz, developed a much more powerful sending mechanism, and used an aerial, or antenna, to radiate the energy of the oscillations more effectively into space. These improvements, coupled a few years later with the use of the Branly coherer as a detector, made wireless telegraphy commercially possible.

A transmitting station used by Marconi in 1896 is shown diagrammatically in Fig. 162. Here a coil of inductance L was inserted between one terminal of the spark gap and the antenna.

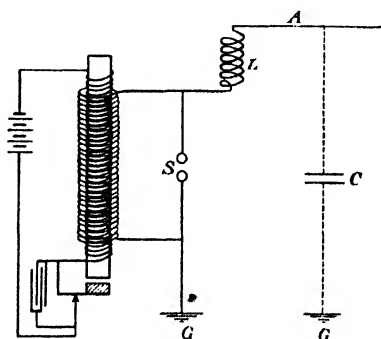


Fig. 162.

The other terminal was grounded, while the antenna and the ground formed two "plates" of an air condenser of capacitance C .

743. Detectors. Among a variety of detectors which have been used during the evolution of radio communication, those that act as valves have been the most successful. Valve detectors are devices which allow a current to pass in one direction, but either stop it altogether, or greatly diminish its magnitude in the other.

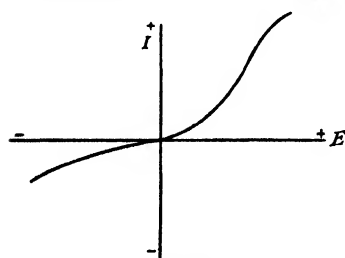


Fig. 163.

The simplest of these is the crystal detector, which consists of a fine needle point pressing upon a crystal such as galena, iron pyrite, bornite, or carborundum. If the point on the crystal at which the contact is made is properly chosen, this device exhibits remarkable resistance asymmetry, yielding

a very high resistance to currents which flow one way, and a very low one to currents in the other direction. This fact is represented

by the characteristic curve in Fig. 163, which shows a large current when a positive e.m.f. is applied to the detector, and a relatively small one when E is negative. The crystal, like all valves, almost

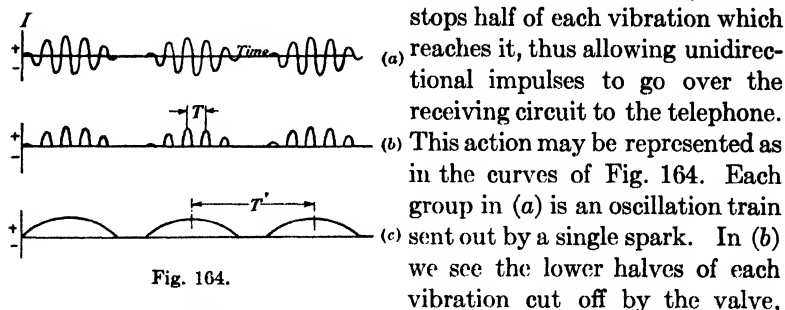


Fig. 164.

stops half of each vibration which reaches it, thus allowing unidirectional impulses to go over the receiving circuit to the telephone. This action may be represented as in the curves of Fig. 164. Each group in (a) is an oscillation train sent out by a single spark. In (b) we see the lower halves of each vibration cut off by the valve, but with the upper halves allowed to pass as a current through the receiving circuit. They each start with small amplitude, reach a maximum, and then decrease to zero. This is typical of a wave train as set up in receiving circuits by a primary oscillation whose first vibration is usually the largest, as in Fig. 156. As neither the diaphragm of a telephone receiver nor the ear can respond to such high frequencies as those of a spark, the result on the diaphragm is a single impulse for each train, as shown in (c). But if these sparks follow each other at a constant time interval T' , the succession of impulses may be at an audible frequency, giving rise to a note of definite pitch. This note is of longer or shorter duration, according to the number of wave trains emitted by the source. The sending operator controls the duration of the tone and produces a short one for the "dot" and a longer one for the "dash" of the telegraphic code.

A much more sensitive valve is the thermionic tube having either two or three "elements." This valve depends upon the fact that if one of two electrodes in an exhausted tube is heated to incandescence, electrons can flow from the heated electrode to the cooler one, but not in the reverse direction. In Fig. 165, the cathode is shown as a metallic filament heated by a current from battery "A", while the "thermionic current" is supplied by the "B" battery.

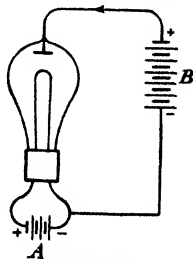


Fig. 165.

In using this valve as a detector, a transformer T (Fig. 166) is connected in the aerial-to-ground circuit, and its secondary terminals are applied to the plate and filament as shown. A variable condenser C is connected across the secondary to permit "tuning" the system to

the frequency of the incoming oscillation. This means giving the receiving circuit the same natural period of vibration as the incoming waves, thus establishing resonance. The currents due to each half wave are then eliminated by the valve, and the telephone receiver R is made to respond to the succession of rectified wave trains, as in the case of the crystal detector.

The explanation of this valve action, and the operation of the much more usual three-element tube, will be discussed in Chapter 56.

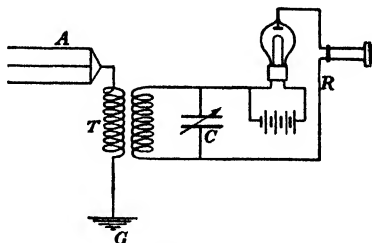


Fig. 166.

744. The Tesla coil. This device, designed by the American inventor Nikola Tesla, is extremely useful for developing high-voltage oscillatory currents. It consists of a step-up transformer, T_1 (Fig. 167), having a ratio of 100:1 or more, and connected to an oscillatory circuit indicated by the curved arrow. This circuit contains a condenser of capacitance C , a spark gap G , and the self-induction L of a dozen turns or so of coil P , which is the primary of a second transformer, T_2 , constructed without iron. Spark discharges across G

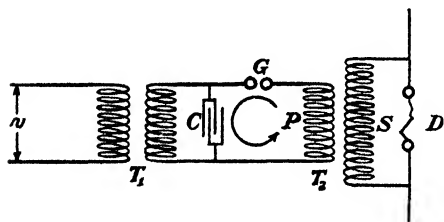


Fig. 167.

oscillate with a frequency determined by the values of C and L . As both are relatively small owing chiefly to the absence of iron in T_2 , the resulting frequency is very high—500 kilocycles or more. The result of these high-frequency currents in

P induces a very high voltage in the coil S , which has many turns of rather fine wire, like the secondary of an induction coil. Thus sparks a foot or more in length and of very high frequency are obtained at D , and make possible a number of striking experiments which depend upon the high frequency of the discharge.

SUPPLEMENTARY READING

C. A. Culver, *Electricity and Magnetism* (Chap. 25), Macmillan, 1930.

A. Zeleny, *Elements of Electricity* (Chap. 27), McGraw-Hill, 1930.

H. Hertz, *Electric Waves*, Macmillan, 1893.

O. F. Brown, *The Elements of Radio-Communication*. Oxford University Press, 1927.

PROBLEMS

1. In an oscillatory circuit in which R is small compared to L , the capacitance is 0.4 microfarad, and the inductance is 16 millihenries. What is the frequency? *Ans.* 1.99 kilocycles.

2. An oscillatory circuit has a negligible resistance, 8 millimicrofarads capacitance, and 1.8 microhenries inductance. What is its wave length? *Ans.* 226 meters.

3. Calculate the critical resistance of the circuit in Problem 1. *Ans.* 400 ohms.

PART VI
CORPUSCULAR PHYSICS

CHAPTER 55

Electrical Discharges

745. Corpuscles. The word *corpuscle* means literally a small body, but it has come to be used by physicists to refer to such elementary particles as electrons, protons, and neutrons. It may also be used to refer to the much larger atoms and molecules and to the less tangible energy quantum called the *photon*.

The great advances in modern physics really date from the study of electrical discharges in exhausted tubes by Hittorf in 1868 and Crookes in 1870, a study that was made possible by an improved technique in the production of high vacua. These discharges were not satisfactorily explained until 1899, when J. J. Thomson concluded that minute particles which he called corpuscles existed within the atom and accounted for the observed phenomena. Thus electrons were discovered, and the corpuscular nature of electricity.

In this book we have so far been concerned mainly with matter and energy in quantity. We have made use of such particles as ions, atoms, and electrons only when they helped explain certain large-scale processes such as electrolysis. But in the "new physics" which follows, corpuscles are studied directly, as well as phenomena which could not be investigated at all without some knowledge of the inner nature of matter and the particles of which it is composed. In these investigations the quantum theory and the concept of the photon have been the basis for a vast array of discoveries concerning the inner structure of the atom.

746. Appearance of the spark discharge. Let a glass tube, equipped with disc electrodes as shown in Fig. 1, be progressively

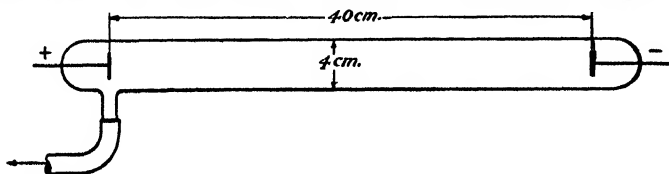


Fig. 1.

exhausted by an air pump. A discharge sent through it from a unidirectional source of high potential assumes an appearance of luminosity which varies as the pressure is reduced. The character-

istic stages of the discharge are the same in all tubes, but occur at different pressures depending upon the tube's dimensions—especially its length. The following data are based on observations of a tube with a gap of 40 cm between the terminals, and a diameter of 4 cm. To produce a discharge between the discs at atmospheric pressure, four or five hundred thousand volts are necessary, and the discharge is an ordinary spark like miniature lightning.

When the pressure of the air in the tube is reduced to 40 mm of mercury, 10,000 volts are sufficient to produce a discharge. This occurs along a sinuous path of much larger section than an ordinary spark, and has a decidedly pinkish color. The general form of this discharge is shown in Fig. 2 (a). At 10 mm pressure (b), the column is thicker, still pinker in color, and has separated from the cathode,

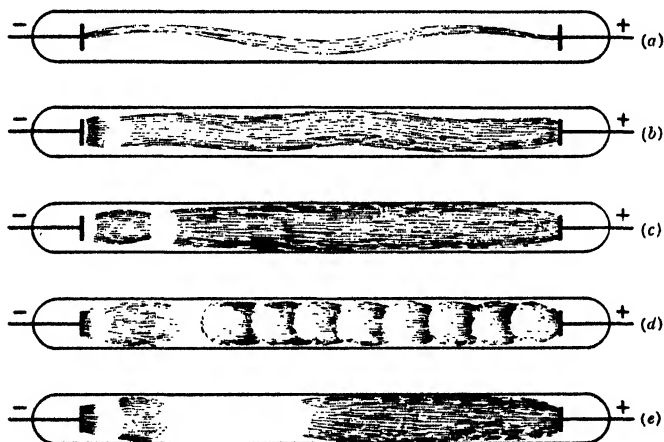


Fig. 2.

at which a violet-colored glow is now visible. This is called the **negative glow**.

At 6 mm, sometimes called a Geissler vacuum because it is about the value used in "Geissler tubes," the **positive column** spreads out to the walls of the tube, becomes paler, and the negative glow advances considerably, as shown in (c). At 3 mm the positive column breaks up into striae, as shown in (d). These luminous bands vary in length and number with the pressure and nature of the gas and current density. They are most marked in impure gases, or mixtures like air, and are most easily obtained in long tubes of small diameter.

The dark space between the positive and negative columns is known

as the **Faraday dark space**. This becomes larger and longer with still higher vacua, and the positive column gradually disappears as a pale white cloud while the negative glow begins to separate from the cathode, leaving a new dark space. The latter, named the **Crookes dark space** after Sir William Crookes, who discovered it,[†] is clearly visible at a pressure of 2 or 3 mm, while a new glow appears on the cathode, spreading gradually over its entire surface. This



Plate 22.

Photograph of discharge through partially exhausted tube, showing Crookes dark space close to the cathode, negative glow, Faraday dark space, and striated positive column. The pressure is about half a millimeter.

“cathode glow” is associated with the ejection of metallic particles from the cathode, whereas the positive and negative columns have to do only with the gas in the tube.

The Crookes dark space gradually enlarges, and at half a millimeter of pressure it is about 2 mm long. At 0.1 mm it extends to 1 cm (Fig. 2 (e)), and at 0.01 mm it fills the tube, the negative column having disappeared. The glass of the tube fluoresces with a greenish light around the Crookes dark space, and this band of fluorescence advances until at pressures around 0.01 mm the whole tube is luminous, though there is no longer any luminosity from the gas it contains.

747. Paschen's law. The potential at which a discharge will just take place between electrodes in a gas was shown by Paschen in 1889 to depend both upon the length l of the gap and on the pressure p of the gas, and he found that E is a function of the product pl . This relation between E and pl is shown by the curve of Fig. 3, where x_0 marks a distinct minimum voltage. In air between parallel-plate electrodes, the minimum voltage, E_0 , equals about 330 volts with pl between 5 and 7, where p is expressed in millimeters of mercury and l in millimeters. Thus if we take 6 as an average value, a dis-

[†] It is also called the **cathode dark space**, and in Germany, the **Hittorf dark space**.

tance of 10 cm between electrodes calls for a pressure of about 0.06 mm, while at 20 cm it would have to be 0.03 mm if 330 volts are

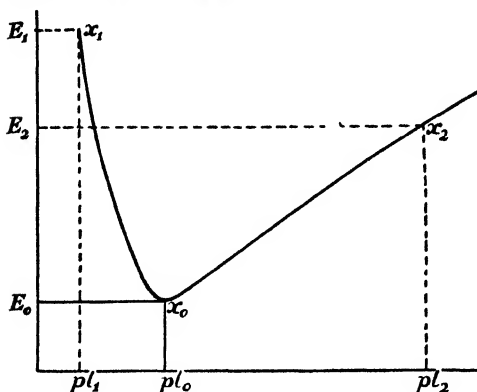


Fig. 3.

to produce a discharge. The amount of a gas, and consequently the number of its molecules which lie between the electrodes in a given tube of uniform cross section, vary as pl ; therefore it seems safe to conclude that a certain definite number of molecules are needed between the terminals to allow the discharge to pass most easily. If there are

either less or more than this optimum number, a higher voltage is necessary.

We may also interpret Paschen's law in terms of the mean free path λ of the particles. Sir J. J. Thomson points out that as λ varies inversely as the pressure, the product pl varies as l/λ . But l/λ is the number of mean free paths between electrodes, and this number must determine the sparking potential. Thomson also calls attention to the fact that λ/pl has nearly the same value of about 0.13 for most gases.

748. Hittorf's experiment. A well-known experiment by Hittorf strikingly illustrates Paschen's law. A discharge is passed between the terminals a and b (Fig. 4) only a few millimeters apart, indicated by l_1 . A longer path, l_2 , is also made available by means of a side tube, as indicated. Since the distances l_1 and l_2 are constant, we may now plot the voltage against pressure instead of pl , and so obtain the two curves for the two paths shown in Fig. 5.

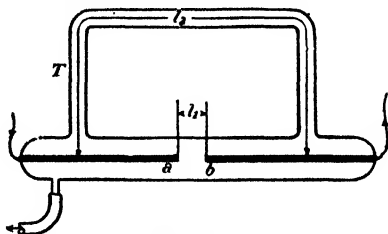
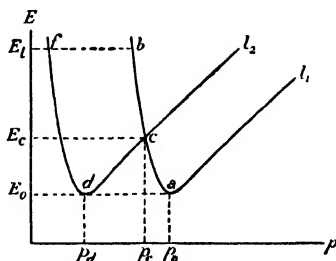


Fig. 4.

If the pressure in the tube is progressively reduced from high values, the discharge, as shown by the right-hand curve, passes more and more easily, until at p_a the curve reaches the minimum a with the voltage E_0 . A still further reduction of pressure demands a

relatively large increase in potential, shown by the steeper slope between a and b . If the source of potential were limited to a maximum value E_i , the discharge would cease at b . But at c the longer path l_2 becomes available and the discharge now passes more easily by this route, with a new minimum at d . At still lower pressures the discharge passes with increasing difficulty and ceases at f when the supposed limit of available potential E_i is reached.



It is interesting to note that the pressure p_c determines a critical and probably unstable condition when either path is available, and that if the voltage is then even slightly below E_c , no discharge can pass by either route. It would pass, however, if the pressure were either raised or lowered.

749. Spark-discharge potentials. The voltage necessary to produce a discharge across a given gap decreases steadily from atmospheric pressure to a certain minimum, different for each gas, and as we have just seen, increases rapidly for still lower pressures, tending to become infinite with a perfect vacuum. That portion of the curve corresponding to high values of pl between parallel plates is approximately a straight line whose equation is $E = A + B(pl)$, where A and B are constants. If l is measured in centimeters, and p in millimeters of mercury, then for air, $A = 1700$ and $B = 39$, so that at any assumed pressure we can readily calculate the necessary voltage for a discharge l centimeters in length. If the pressure is one atmosphere, the slope of the curve is about 31,000 volts per centimeter between parallel plates whose edges are shielded. This value, because of the increasing field concentration with increasing curvature, becomes much smaller if the terminals are curved. Between points, the discharge occurs at still lower potentials. The critical potential gradient is not constant with increasing distance between curved electrodes. Instead it tends to decrease with increasing distance. This means that spherical electrodes, as they are more and more separated, behave increasingly like sharp points.

Between polished brass balls 2 cm in diameter, the critical potential is 31,000 volts with a spark gap of 1 cm. With a gap of 2 cm it is 23,000 volts per cm. With a gap of 4 cm it is only 16,000 volts per cm. Values for much longer discharges have been obtained in the high-

tension laboratory of the General Electric Company at Pittsfield. Between points the potential gradient was found to be nearly constant, with increasing spark length up to a distance of nearly 5 m in air at atmospheric pressure. This value is about 4000 volts per centimeter, or 10,000 volts per inch. The same figure also holds approximately for discharges between small parallel wires.



Courtesy General Electric Co.

Plate 23.

Thirty-foot spark of ten-million-volt artificial lightning discharge in high-voltage engineering laboratory, Pittsfield, Mass. Energy output, 125,000 watt seconds.

750. Cathode rays. In 1868, Hittorf, a German physicist, while studying the discharge in exhausted tubes, observed the dark space between the cathode and the negative column, and was led to a study of what he considered to be some form of wave motion which spanned the gap, and caused the gas to glow beyond it. In 1869, by producing still higher vacua, he nearly eliminated the glow, and found instead that the whole tube fluoresced, while obstacles in the path of the rays cast a shadow on the part of the glass that was screened from their action. A year later, Crookes showed that these rays could set a small mill wheel, having light mica vanes, in rapid rotation, and concluded that the rays were particles of a new kind of matter lying in the borderland between matter and energy. In this view he was sustained by other British physicists, including Maxwell. In 1895

Jean Perrin, in Paris, showed that this cathode-ray material was negatively charged. But it was not until 1899 that J. J. Thomson, of the University of Cambridge, announced the conclusion, based on his own experiments, that the cathode rays were "corpuscles" contained within the atom, and that these corpuscles were ejected from the atoms of the cathode during the discharge.

The charge e on these particles, now called electrons, had been measured in 1897 by Townsend at Oxford, who assigned to it the value of -3×10^{-10} esu, and in 1898 and 1899 by Thomson himself, who obtained results more than twice as large. Thomson's value, coupled with measurements both he and others had made of the ratio of the charge to the mass, e/m , of an electron, made possible a calculation of m , which Thomson estimated at 3×10^{-28} g. Probably the best present value is 9.11×10^{-28} g. It is obtained from $e/m = 5.272 \times 10^{17}$ esu/g and $e = 4.803 \times 10^{-10}$ esu. These values are based on X-ray spectroscopic measurements, as well as on Millikan's oil-drop experiment, described in the next article.

The ratio of the mass of the hydrogen ion (proton) to the mass of the electron may be found, if we make the reasonable assumption that they both have the same elementary charge, one positive, the other negative. In Article 650 we saw that the ratio e/m_H equals F numerically, where F is 96,494 coulombs. This is 9649.4 emu/g, and multiplying by 2.998×10^{10} (the velocity of light), we have $e/m_H = 2.8929 \times 10^{14}$ esu/g. Now if we divide $e/m = 5.272 \times 10^{17}$ esu/g (the value for the electron given above) by e/m_H , we obtain $m_H/m = 1822$. Thus, using these data, we find that the hydrogen ion is 1822 times heavier than the electron. This figure, however, is too small, and 1835, obtained by spectroscopic measurement, is considered a better value. The mass of the hydrogen ion may now be calculated by taking the product $1835 \times m$, which equals 1.672×10^{-24} g.

751. Millikan's measurement of the electron charge. Our feeling of certainty regarding the electron theory has been greatly strengthened by the researches of Professor Millikan, who, during the period from 1909 to 1917, at the University of Chicago, made numerous measurements of the charge e by a very different method from those employed by previous observers, and found $e = -4.774 \times 10^{-10}$ esu, a figure which is slightly smaller than the more recent value quoted above. These experiments also verified the assumption that this is the ultimate charge of which all ionic charges are composed, for no ions having a smaller charge, nor ions having charges not an integral multiple of e , have ever been observed.

Millikan's method consisted essentially in observing the behavior of minute droplets of oil in an electrostatic field after they had captured one or more electrons. The oil was blown through the nozzle *N* from the atomizer *A* (Fig. 6), and a drop such as *D* was illuminated

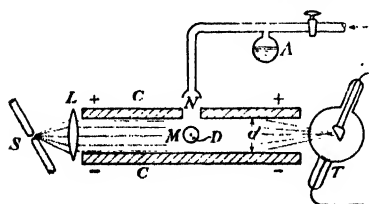


Fig. 6.

by light from an arc *S* concentrated by the lens *L*. This made the drop visible to an observer looking through a microscope directed normally to the diagram, as indicated by the circle *M*, which represents its objective. The droplet then appears like a tiny star against

a dark background, as when the Brownian particles are observed through an ultramicroscope.

In the diagram, *CC* represents the sections of condenser plates 1.5 cm apart and maintained at a measured potential difference of several thousand volts. Some of the drops, as a result of friction in the atomizer, become charged with one or more electrons. In a field of suitable intensity, a charged drop may be supported against gravity and seen at rest in the field of the microscope. But under the influence of X-rays from the tube *T*, the charge on the drop may either increase or decrease, causing it to move up or down. It may then be again balanced by changing the field intensity. When in equilibrium, the drop's weight mg equals the field intensity E (or V/d), times the charge. The equation is

$$mg = \frac{Vq}{d}. \quad (1)$$

The mass m was found by an application of a law due to Stokes, which makes it possible to calculate the radius of a small drop as it falls through a gas in terms of its speed, which is slow in the case of the oil drops observed. Then with the radius and density of the drop known, its mass m is easily calculated. In Millikan's experiments, q was always an integral multiple of an elementary charge e , or $q = ne$ where n is a whole number. The smallest value of q is of course equal to e , when $n = 1$. Then if V and d are accurately measured, e may be calculated from equation (1).

752. Velocity of cathode rays. The stream of negative charges that constitutes the cathode rays is equivalent to a current flowing in a flexible wire. Therefore, cathode rays are deviated by a transverse magnetic field. The force acting upon them may be obtained from the equation $F = BIl$ (Article 731). But any current I may be

measured in terms of the time rate of flow of a charge q . If I is uniform, $I = q/t$, and if we consider a current element of length l , then $Il = ql/t$. But if q is the electronic charge e , l/t represents its veloc-



Plate 24.

Pencil of cathode rays, made visible by fluorescent screen, and curved upward by a magnetic field directed toward the observer.

ity. Therefore, $Il = ev$, and the force on this equivalent current, acting at right angles to the direction of the electron stream, is given by

$$F = BIl = Bev. \quad (1)$$

This force is evidently constant for a given velocity, provided B is uniform over the range of motion considered. Such a constant force acting perpendicularly to the velocity is the necessary and sufficient condition for circular motion; therefore the path is an arc of a circle lying in a plane at right angles to the field.

In Fig. 7 the electron path is shown entering a uniform magnetic field at a . This field is normal to the paper, directed away from the observer, and supposed uniform between AA' and BB' .

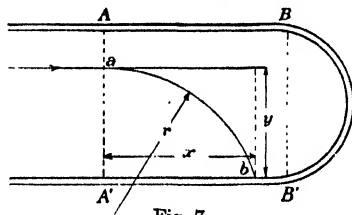


Fig. 7.

From a to b , where the electron meets the wall of the tube, the path is an arc of a circle of radius r . But the force required to hold a body of mass m in such an orbit is given by $F = mv^2/r$; hence, equating this value of F with that of equation (1), we obtain

$$Bev = \frac{mv^2}{r}.$$

$$\therefore v = \frac{Bre}{m}. \quad (2)$$

753. Measurements with crossed fields. The preceding experiment alone is not sufficient to determine v , because it depends upon

the value of the ratio of the charge to the mass of the electron. Among various ways of measuring e/m , the method of crossed electrostatic and electromagnetic fields has proved especially useful, as with the same apparatus we may find both v and e/m .

In Fig. 8 is shown a tube in which an electrostatic field of intensity E is crossed with the magnetic field just discussed. The cathode rays from C pass through a hole in the anode A and then through another hole in the diaphragm D so as to form a narrow pencil of rays.

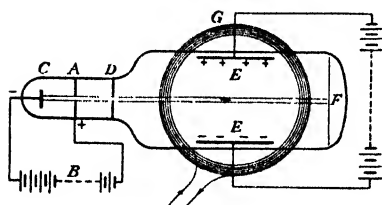


Fig. 8.

This pencil strikes a fluorescent screen F , where it produces a luminous spot which moves under the influence of the fields. When the pencil of rays passes between the charged plates EE , it is acted on by an electrostatic field which exerts a constant upward force in the case illustrated. This would

produce a parabolic path like that of projectiles acted on by gravity alone. But if the magnetic field produced by the coils G is properly adjusted, the two effects may be made to neutralize each other, and the beam is undeviated. Then the electrostatic force eE is equal and opposite to the magnetic force Bev , or $eE = Bev$, and $v = E/B$. Thus, we may find the velocity without knowing the value of e/m .

The measurement of e/m is made by first sending the electron stream through an accelerating field of known value. This is set up between the cathode C and anode A , by a battery B . The kinetic energy acquired by the electron during the time of its acceleration is equal to the work done upon it, which is the product of the charge e and the potential difference between C and A . If this potential is represented by V , we may write $Ve = \frac{1}{2}mv^2$. But when the fields are balanced, $v = E/B$; hence $\frac{1}{2}mE^2/B^2 = Ve$, and

$$\frac{e}{m} = \frac{E^2}{2B^2V}, \quad (1)$$

from which the ratio is determined.

754. Variation of mass with velocity. Equation (1) above shows the possibility of measuring the mass of an electron from a determination of the ratio of the charge to the mass, and a knowledge of the charge e . In 1901 and 1902, Kaufmann made a series of observations on the mass of electrons moving at different speeds, and found that it

varies with the speed. Up to about half the velocity of light it remains practically constant at the accepted value, but beyond that velocity, the mass increases more and more rapidly, tending toward infinity as it approaches the velocity of light, following a curve like that in Fig. 9. The theoretical equation of this curve, based on the principles of relativity, is $m = m_0(1 - \beta^2)^{-1/2}$, where m_0 is the mass of the electron at rest, and β is the ratio of its velocity to the velocity of light

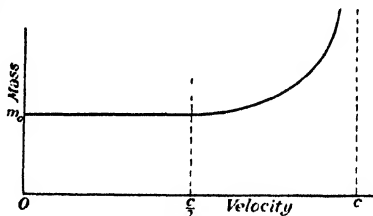


Fig. 9.

c . When $\beta = 0.5$, the calculated value of m is $1.155 m_0$; when $\beta = 0.9$, it is $2.29 m_0$; for $\beta = 0.99$, $m = 7.089 m_0$, and when $\beta = 0.999$, $m = 22.36 m_0$.

755. Other characteristics of cathode rays. In addition to being deviable in electrostatic and electromagnetic fields, cathode rays have certain other remarkable properties. When not deviated by a field, they move in straight lines, and unless influenced by the shape of the tube, these lines are normal to the surface of the cathode. The position of the anode has no effect upon their direction, so that it may be located on one side of the tube without influencing the trajectory of the electrons. If carried outside the tube through a "window" made of thin aluminum foil, they are found to possess powerful ionizing properties, rendering air and other gases conductive. They make profound chemical changes in substances upon which they strike, as has been shown by Coolidge in the General Electric Research Laboratories. They cause certain crystals to become fluorescent, as well as the glass of the tube in which they are produced, and they heat any obstacle that comes in their path. In fact, the "target" in X-ray tubes may become incandescent under their bombardment. This is not surprising, considering their high velocity. A speck of dust, just visible to the naked eye, weighing 0.1 mg and moving with one tenth of the velocity of light, would have kinetic energy equal to that of a mass of three tons after falling a mile! The electron is vastly smaller than this speck of dust, but when present in such numbers as in a cathode beam, the aggregate is very impressive.

The energy of moving electrons is measured in *electron volts*, which means the energy imparted to a charge e when it moves without collision through a rising potential of V volts. One electron volt equals 1.59×10^{-12} erg, but as fields of over a million volts

are now possible, an electron may be given kinetic energy above 2×10^{-6} erg.

756. The cathode ray oscillograph. Because of its almost negligible inertia, a pencil of cathode rays is an ideal indicator of a rapidly changing magnetic or electrostatic field. In commercial studies of alternating-current wave forms and of electromagnetic oscillations in general, highly exhausted tubes are used in which a willemite screen, *W*, fluoresces at the spot where the pencil *p* of moving electrons strikes it, as indicated in Fig. 10. The cathode is heated by a device not shown, causing it to emit electrons at a relatively low voltage. The cathode beam is reduced to a narrow pencil by the ring *R* (an auxiliary

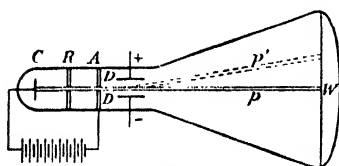


Fig. 10.

anode in some recent types) and by the perforated anode *A*. It then passes between two pairs of parallel plates, only one of which, *DD*, is shown. This pair, when charged to a potential difference as indicated, deflects the pencil upward, as is seen by the displacement of the luminous spot on the screen. If the charge is reversed, it is bent downward, while right and left deflections are produced by the other pair of plates. By combining these motions, the spot may be made to trace out Lissajou's figures from two oscillatory sources of different frequencies, or it may be made to give a hysteresis loop.

If the horizontal deflecting plates are connected to a voltage source that builds up gradually to a maximum and then drops suddenly back to zero, the beam may be made to move periodically across the screen at a uniform speed and then snap back to the starting point. A circuit which will provide such a voltage is called a **sweep circuit**. It depends in part upon the remarkable ability of the thyatron (Article 769) to start and stop a current abruptly at definite potentials. With this device the vertical deflection due to an alternating field impressed upon *DD* will cause the spot to trace out a curve against time. This curve is really a snapshot of successive cycles held stationary by the sweep circuit, and thus exhibits the "wave form" of the alternating e.m.f.

757. Discharges and ions. In order that a gas may conduct a current, ions must be present in it, as when liquids conduct. Negative ions are either electrons, or larger particles as atoms, molecules, or groups of molecules having one or more excess electrons attached to them. Positive ions are atoms or molecules or molecular groups each

deprived of one or more electrons. A few free ions exist normally in the atmosphere, but not enough to make it conduct appreciably. The number present may be greatly increased by X-rays, rays from radium, and so forth. Then the air conducts readily, as is shown by the rapid collapse of the leaves of a charged electroscope. If the gas in an ordinary discharge tube is ionized by X-rays, a measurable current may be sent through it by voltages much lower than is required for the spark discharge. The current increases, as shown in Fig. 11, up to a value I_s , called the **saturation current**, which depends upon the number of ions being formed per second. At this value all the ions created are used before they have a chance to recombine.

Above the potential needed for saturation by a given source of ions, the current stays constant until at say 150 volts, the negative ions move so fast that they create more ions by collision with neutral atoms or molecules. These again create still more, and the current increases again at an accelerated rate with



Fig. 11.

rising potential. After it has reached a value high enough to give the slower-moving positive ions sufficient energy to ionize in their turn, the current develops into a spark discharge.

758. Ionization in a self-sustained discharge. The process of ionization by collision is not limited to gases ionized by an external agency. It takes place in the discharges described in Article 746. The electrons emitted by the cathode first traverse the Crookes dark space, whose length is comparable to their mean free path. With minimum discharge currents, the rise of potential across it varies from 400 down to 60 volts, according to the nature of the cathode. This is the major part of the total rise of potential from cathode to anode, and the field intensity there is so very great that the electrons emitted acquire a high velocity. They are then able both to ionize the gas and to make it luminous in the negative glow just beyond the dark space. Different gases become luminous with different energies of the bombarding electrons, as is indicated by the different "resonance potentials" required. It takes 19.75 electron volts to excite helium to luminosity, but only 4.9 electron volts to excite mercury vapor.

At a still higher voltage the impact of the electron is sufficient to knock out an electron from the gas atom. This requires 24.5 volts with helium, and 10.4 with mercury vapor. The result is a new electron that travels on toward the anode along with the first electron, and also a positively ionized atom which moves toward the cathode. If there are many electrons in the stream, ionization occurs simultaneously with luminosity at resonance potential, for then successive collisions with the same atom in its excited or light-emitting state may result in knocking out an electron and so ionizing it. When this occurs the current rises very abruptly after the resonance potential has been reached.

The positive ions created by collision move much more slowly than the electrons, and are more feeble ionizers, even if we assume equal kinetic energies.

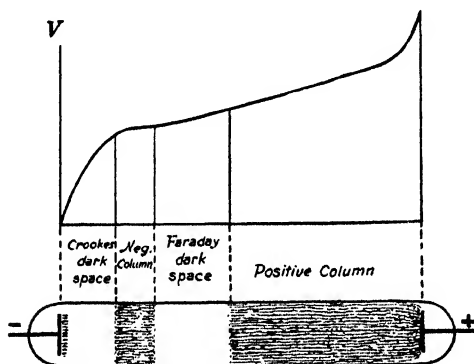


Fig. 12.

However, if the voltage is raised above the ionizing potential, they begin to ionize in their turn, and the current is still further increased, tending toward an *arc discharge*, which will be discussed later.

In the negative glow, the potential does not change much, but begins to rise again across

the Faraday dark space. This is a second region of acceleration for the electrons that have been slowed down by collisions in the negative glow. In the positive column the potential rises at a gradual and nearly uniform rate, except when the striae appear, but there is a sharp increase close to the anode. Its luminosity is due to both kinds of ion which *recombine* along its entire length. These variations in potential are shown approximately in Fig. 12.

759. Origin of cathode rays. In tubes such as we have described, having a gas pressure of say 0.01 mm of mercury, it is found that if the tube is bent down in front of the cathode, the rays develop only from that portion of its disc not so shielded. Thus the depression *B* in the wall of the tube shown in Fig. 13 prevents the rays from forming over the upper portion of the cathode. The explanation is that positive ions derived from the gas in the tube move toward the cathode

under the influence of the field, and liberate electrons by their impact with the cathode. Thus ions are needed to liberate the very electrons which are to form them.

The velocity required to enable positive ions to knock electrons out of the cathode is developed in the strong field of the Crookes dark space. The total change of potential there, as we have seen, has a minimum value of from 60 to 400 volts, whereas only 25 or 30 volts are needed to give the electron enough energy to ionize the gas in the negative glow. It would seem then that the large potential drop across the dark space is determined by the needs of the relatively heavy and slow-moving positive ions.

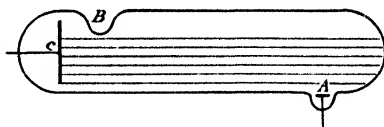


Fig. 13.

760. Canal rays. The positive ions just referred to have their origin in the gas of the tube between anode and cathode. If the cathode is pierced with small holes (or canals), streams of these ions pass through, forming what Goldstein, who first observed them in

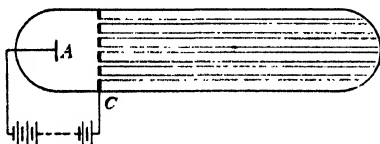


Fig. 14.

1886, called **canal rays**. These rays, shown in Fig. 14, appear as a yellowish glow in air at low pressure behind the cathode, and travel from 10 to 100 times slower than cathode rays. They are deviable in

both electrostatic and electromagnetic fields, but not nearly as much so as cathode rays because, although their slower speed should make them easier to deviate, positive ions are at least 1825 times more massive, and so their greater inertia resists deflection. Equation (2) of Article 752 applies to canal rays as well as to cathode rays, and the radius of curvature in a magnetic field is given by $r = (v/B) \times (m/e)$, where m is the mass of the individual particle. The mass is obtained by measuring the radius in a known field. The results show that these rays are made up of protons, as well as ions of the gases in the tube, and even of the metal of the cathode.

In addition to true canal rays, the anode itself, when heated, gives off positive ions of its own substance or of metallic salts with which it may be coated. This fact has been used in some brilliant investigations by J. J. Thomson, Aston, and others in order to determine atomic masses. The methods used by Aston and several other physicists are discussed in Articles 789 and 790.

PROBLEMS

1. Using the formula in Article 749, calculate the critical discharge voltage between parallel plates 12 cm apart, when the air pressure is one eighth of an atmosphere. *Ans.* 46,160 volts.

2. In a transverse field of 150 oersteds, a cathode beam traces a curve of 6 cm radius. Taking the value of e/m in Article 750 (converted to e.m.u.), calculate the velocity of the electrons. (NOTE: an emu = 3×10^{10} esu.) *Ans.* 1.58×10^{10} cm/sec.

*3. Calculate the electrostatic field needed to balance a magnetic field of 200 oersteds in the "crossed fields" experiment, when the accelerating e.m.f. is 90 volts. *Ans.* 1125 volts/cm.

CHAPTER 56

Thermo- and Photoelectric Emission

761. Thermionic emission. The valve used in wireless telegraphy and described in Article 743 depends upon a phenomenon first really studied by Elster and Geitel about 1880, but not thoroughly understood until the exhaustive investigations of Richardson in 1900 and after. When a metal is heated to incandescence, it tends to lose negative electricity, and so, if charged negatively, it rapidly loses the charge. Positive electricity may also be discharged in this way, but bodies showing this anomalous behavior, which is probably due to impurities, do not retain it upon prolonged heating. Thus true thermionic emission is a property of certain metals which does not alter with time, and occurs whenever they are sufficiently heated.

Edison observed this phenomenon in connection with incandescent lamps before 1885, and found that if an electrode were sealed into the lamp independent of the filament, a current could pass from the electrode to the filament, but not in the opposite direction. We now know that what occurs is an emission of electrons from the heated body. In a vacuum under a sufficient potential gradient they move with high velocities, and constitute cathode rays like those already described. Their emission may be facilitated by a deposit of lime or thorium oxide on the cathode, which causes the emission to take place at a lower temperature, and a lower voltage may then be used in producing the cathode rays.

This method of production, with a heated cathode, differs from the usual way in a very important respect. It is wholly independent of the gas that always remains in an exhausted tube, and so works with any degree of vacuum. In fact, the discharge is readily created when it would be quite impossible with a cold cathode, because then we should be dependent upon the positive ions in the gas, and such ions would not be present in sufficient quantity to liberate electrons by their impact.

Richardson compared the emission of electrons from a heated body to ordinary evaporation, and actually calculated the "heat of evaporation of electricity," which is the energy required to liberate a unit quantity from that body. As the liberation of a negative charge must

leave a corresponding positive charge on a conductor, a "double layer" is formed at the surface, as occurs when an electrode is immersed in an electrolyte. To overcome this adverse field, a minimum potential, which differs for different metals, is necessary to eject the electrons from its influence. It is found to be 4.1 volts for platinum, 3.04 for heated lime, and 6.48 for carbon, which shows why a spot of lime on a heated strip of platinum is such a prolific source of electrons, and why both lime and platinum are better than carbon.

762. Thermionic currents. If the temperature of the heated body is kept constant, increasing voltage results in increasing current, but the increase, unlike metallic conduction, is not steady. Instead, the current-potential curve is as shown in Fig. 15. The current increases

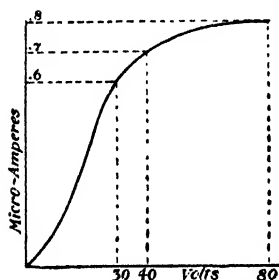


Fig. 15.

rapidly at first, then more and more slowly as it begins to make complete use of all electrons emitted by the cathode. This decline in its growth begins at the "knee" of the curve, or at about 0.6 microamperes in the particular case illustrated. After the "knee," the current approaches a saturation value when all the electrons liberated in a given time reach the anode.

The principal factor which controls the thermionic current is the so-called **space charge**. This is really the combined charge of all the electrons between the electrodes. The total effect of these many minute charges is as if they formed a charged cloud which repels the advance of other electrons from the cathode. Of course, the electrons that form this "cloud" are in rapid motion, but as there are always new ones in the same space, the space charge does not move. The thermionic current would reach saturation more readily than it does if there were no space charge. That is, a lower voltage would produce saturation. Below the knee, the space charge is the chief factor controlling the current. Beyond the knee, the space charge is less effective, owing to a thinning out of the electrons by their rapid motion, and the current is mainly limited by the total number of electrons emitted by the cathode.

Since the value of the saturation current is limited by the rate at which the cathode emits electrons, this limit rises if the cathode is heated to a higher temperature. These relations are shown by the curves of Fig. 16, where the curve *OA* is the first part of the volt-ampere characteristic, if we assume an unlimited supply of electrons

from the cathode. The curve Ot_1 shows a definite limit I_1 , reached when the cathode is kept at a temperature t_1 , while the curves Ot_2 and Ot_3 show higher limiting values of the current at temperatures rising in equal steps. The equation of the curve OA , due to Langmuir, is $I = kV^{\frac{2}{3}}$, where k is a constant depending upon the construction of the tube.

There are three ways of varying the current in a thermionic tube having constant filament temperature. One, as we have seen, is by varying the plate potential. A second is by altering the space charge by means of a "grid" (Article 763). The third is by suppressing the space charge by the use of an inert gas. This method is used in the "tungar" rectifier, which is used to "rectify" alternating currents for charging storage batteries and for other purposes where half of every cycle must be suppressed. In this tube the cathode is a heated tungsten filament, and the anode is a tungsten cone a few millimeters from the cathode. Instead of being highly exhausted, it contains argon gas at a pressure of from 8 to 10 cm of mercury. This gas supplies positive ions under the bombardment of the cathode discharge, and thus neutralizes the space charge and permits the current to increase almost indefinitely under increasing voltage, without reaching a saturation value.

763. The three-element tube. In this form of thermionic tube a third electrode is introduced into the thermionic tube between anode and cathode, known as the **grid**. The cathode filament is raised to

incandescence by the "A" battery, and thus emits electrons in great quantities. The tube is so highly exhausted that the effect of the space charge in limiting the current is very marked, even with the rather low potentials used. A conventional diagram of such a tube used in a radio receiving circuit is given in Fig. 17.

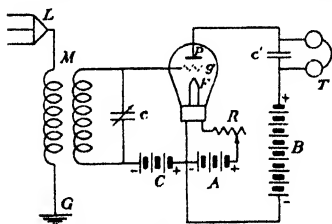


Fig. 17.

The "A" battery, controlled by a rheostat R , heats the filament F . The "B" battery of 40 volts or more furnishes the so-called plate current, which operates the head phones, T . This current flows from the anode, or plate P , through the tube of the filament F , because negative electrons are moving from F to P . The grid g is

connected to the receiving circuit. This is "coupled" by means of the coils M to the aerial, or antenna, L , which is grounded at G . A condenser c of variable capacitance is connected across the secondary of M so as to tune the receiving circuit to resonance with the currents set up in the aerial by the incoming radio waves. Finally, a "C" battery of a few volts may be inserted in the receiving circuit so as to give the grid a fixed difference of potential, usually negative, with respect to the filament.

To understand the operation of the grid, let us suppose that there is no "C" battery, and that it is given a small positive charge from the receiving circuit. This tends to neutralize the space charge, and permits an increase in the plate current. When the receiving circuit charges it negatively, it acts like an increased space charge, and the plate current is reduced.

The characteristic curve in Fig. 18 shows that small changes in grid potential may produce relatively large changes in the plate current, especially where the curve is steepest.

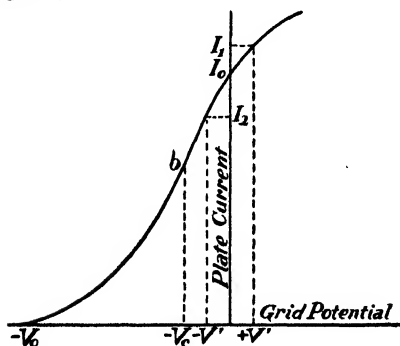


Fig. 18.

Thus if the grid potential is raised from zero to $+V'$, the current increases from I_0 to I_1 , while an equal fall of potential to $-V'$ produces a similar decrease in the plate current to I_2 .

The portion of the curve lying to the left of the current axis shows that if a potential of $-V_0$ volts is given to the grid, the effective space charge is sufficient to reduce the current to zero.

The point b on the curve is a point of inflection where it is steepest and most sensitive. If the grid is "biased" by giving it a small steady negative potential $-V_c$ by a "C" battery, the operative range of the tube may be brought to this portion of the curve. Here also it is straight, so that the current varies directly as the grid potential. There is therefore no *distortion* of the plate current as there would be with the grid at zero potential. With no grid bias, the curve shown in Fig. 18 is convex upward where it is operative. This results in reducing the increase of the plate current with rising grid potential, and in augmenting its decrease with falling grid potential.

764. Use of the three-element tube in radio reception. The three-element tube, or triode, as it is often called, may be used both as a de-

tector and as an amplifier of high-frequency electric oscillations. When used as a detector, the operative range is brought to one of the two most curved portions of the characteristic, and thus deliberately distorts the plate current.

The transformation of small alternating grid potentials into pulsating unidirectional plate currents is shown by the curves of Fig. 19. The sine curve marked *A* represents the variations of the grid potential, and the curve *B* represents the resulting variations of the plate current. These are larger above the undisturbed value I_0 than below it. Thus the average change in current due to the variations *A* is $I_1 - I_0$, where I_1 is the average value of the entire plate current measured from the *X* axis. As the receiver's diaphragm, because of its inertia, is unable to respond to the extremely rapid alternations indicated, the average change of current, $I_1 - I_0$, can cause only a steady pull upon it, and no sound is produced. Even if there were, it would be far beyond the range of audibility.

In order to make the wave train produce audible sound of a definite pitch, the amplitude of the oscillations must themselves have a pulsating change whose frequency is within the range of audibility.

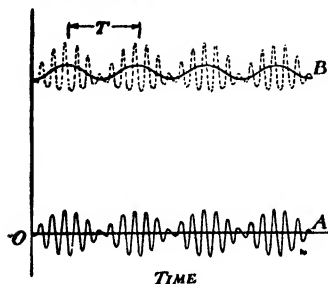


Fig. 20.

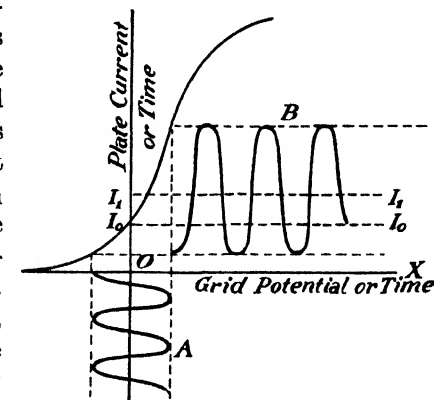


Fig. 19.

They are then said to be modulated. Thus in Fig. 20, the sine curves *A* of varying amplitudes represent the modulated grid potential, and the dotted curves of *B* represent the resulting plate current. This flows in one direction, but fluctuates asymmetrically so that the average is a "rider" of period *T* having *audio frequency*, as distinguished from the rapid alternations of *radio frequency*, which act as "carrier waves."

The carrier wave must be of much higher frequency than that of audible sound, because the energy radiated from an antenna varies as the square of the frequency, and consequently the low frequency

oscillations corresponding to audible sound would radiate very little. Another reason for the carrier wave is that it has a constant frequency, and both the sending and receiving circuits can be tuned to resonate with it. With the variable audio frequencies which must be broadcast, tuning would be impossible. On account of the high frequency of the carrier wave, the telephone receiver T , shown in Fig. 17, would practically stop the current because of its large inductance, which resists rapid current changes. To counteract this effect, a condenser c' is shunted across the receiver. Condensers are increasingly "transparent" as the frequency rises, and the system $c'T$ is thus made capable of responding to the modulated oscillations.

When used as an amplifier, the tube is highly exhausted, and the purpose is to produce audio-frequency plate currents of exactly the same form as the varying grid potential which causes them. These are said to be amplified because the plate potentials needed to produce them (unaided by the grid) would have to be many times greater than the grid potentials actually used. To secure undistorted as well as maximum amplification, the tubes are made to operate at the straight portion of the characteristic curve, as was explained in the preceding article. If this lies to the left of the current axis, as in Fig. 18, a negative "bias" is necessary. The negative charge on the grid is desirable in any case, as it prevents the grid from attracting electrons and so setting up a grid-to-filament current which is harmful. Indeed, "power tubes" used in radio sets after several stages of amplification require quite large biases. In such tubes the grid potential varies through several volts, and if there were not a correspondingly large bias, the grid would become periodically positive, and a periodic current would flow between it and the filament.

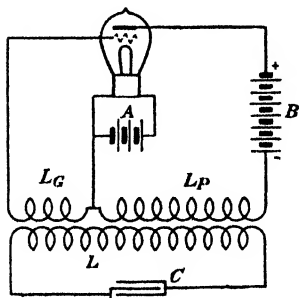


Fig. 21.

765. The triode as a source of oscillations. In order to generate a continuous train of high-frequency oscillations, a condenser and inductance are connected so as to form a resonating circuit, as shown in Fig. 21. This represents a fundamental arrangement from which there are numerous derivatives. An oscillation is started in the circuit LC by

some slight change in the plate current flowing through the coil L_P . This acts inductively on the coil L to induce an e.m.f. sufficient to give C a slight charge. The condenser then discharges, and so oscil-

lations begin, whose frequency is given approximately by $n = 1/(2\pi\sqrt{LC})$, as explained in Article 740. These, in turn, act inductively on coil L_G , and set up corresponding variations in the grid potential, which cause larger variations in the plate current than those which started the process. If the coils are properly wound, the oscillations in the resonating circuit are in synchronism with those of the plate circuit, and build up to a steady maximum value at which they are maintained at the expense of energy derived from the "B" battery.

Such a process, which depends upon making effects help causes, is contrary to the fundamental principle of reaction opposing action. It is possible, therefore, only when energy is constantly supplied (as from the "B" battery) so as to further the action which periodically releases that energy. Thus the original current change in the plate circuit must result in a grid potential which tends to increase that change. The change may be either an increase or decrease of the current, just as the escapement of a clock, at the expense of the potential energy of the weights, causes the pendulum to be driven in that direction in which it is momentarily swinging. The circuit LC acts like the pendulum in maintaining a constant frequency, and also serves to pass the energy of the battery on to the grid.

766. The arc discharge. In the discharges hitherto described, the current is carried by gaseous ions, or by electrons ejected from the cathode, or by both. If the cathode is cold, the electrons are due to the impact of positive ions, but if it is heated, no positive ions are necessary, and the discharge can take place in a high vacuum.

The electric arc differs from these modes of discharge in that the atoms of the terminals play a part as ions in carrying the current. This is strikingly evident in the light of the arc between iron electrodes, which is very rich in the spectral lines of iron. Carbon terminals having a core of graphite mixed with sodium silicate and the salts of boron, calcium, magnesium, and so on, produce a "flaming arc" whose spectrum is that of the elements used in the core.

Another characteristic which distinguishes the arc from the spark discharge is that one or both electrodes are heated to incandescence. The anode may be kept cool, but the cathode must be hot enough to emit electrons, as in ordinary thermionic emission. It is heated by the bombardment of positive ions and not by an auxiliary "A" battery. Thus the arc, even in air at atmospheric pressure, is self-maintained at a voltage far below that required to maintain a spark discharge in an exhausted tube. The carbon arc, for instance, requires only 40 volts after it has once been started, but the start must

be made by touching the carbons together and then "drawing the arc" as they are separated.

The electron stream of an arc, produced from the heated cathode, in turn bombards the anode, and as shown in Article 639, this may produce very intense heat and light from the crater, provided the anode is not cooled. If it is cooled, as by circulating water, the discharge itself may still be produced, but if the cathode is cooled, or is so massive that heat is rapidly conducted away from the region bombarded by positive ions, electrons are not emitted and the arc cannot be formed. In fact, a carbon arc made of one slender and one thick carbon rod acts as a valve in which only the slender rod can act as cathode.

767. Photoelectric phenomena. We may liberate electrons from a metallic surface by the agency of a beam of light, as well as by heating the metal. If light, especially ultraviolet, falls upon a metallic surface, it tends to liberate electrons from the metal, leaving it positively charged. The various metals differ strongly among each other in this respect, and may be arranged in a photoelectric series similar to the voltaic series already given (Article 658) in our study of contact differences of potential. In this series the metals occur in essentially the same order as in the voltaic series. Thus, when illuminated with light of the same wave length, zinc loses electrons much more rapidly than copper, while the alkali metals like potassium are still more photoactive and emit electrons under the relatively long waves emitted by an oil lamp. Rubidium is so extremely active that it loses a negative charge when illuminated by a glass rod barely red hot.

Gases are photoelectric as well as metals, but most of them require light of shorter wave length to ionize them. Air, for instance, cannot

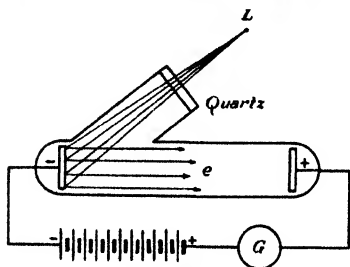


Fig. 22.

be ionized by light that passes through quartz, although quartz is transparent to the near ultraviolet. But the still shorter waves transmitted by fluorite are able to ionize air.

A cell made with two electrodes, one of which is photoactive, and with both sealed in a vacuum tube, acts like a weak battery.

The more photosensitive electrode (for example, potassium), when illuminated as shown in Fig. 22, emits electrons and acquires a positive

charge. If it is connected to the other electrode through an external circuit, a feeble current flows through the connecting wire from the more active to the less active electrode. This means that there is a photoelectromotive force *within* the cell directed from the less toward the more active metal. A battery placed in the circuit, as shown above, may stop the current completely if its e.m.f. is directed so as to oppose the flow of electrons indicated by the arrows, or it may greatly increase the current if connected as shown in the diagram.

The time required to develop photoelectric emission is extremely short, being under 3×10^{-9} sec. both in beginning, and in stopping when the light is turned off. This renders the phenomenon available for many practical applications, such as in television and sound films, through the use of photoelectric cells.

Photoelectrons are ejected with speeds averaging around 5×10^7 cm/sec., and behave exactly like cathode rays. The amount of negative electricity thus liberated varies directly with the intensity of the incident light, but the *velocity* of the ejected electrons depends only upon the wave length of the light, and increases as the wave length decreases.

This latter fact cannot be explained on the basis of ordinary mechanics, and shows that the kinetic energy of the "rays" is derived, not from the interior of the atom, as was formerly supposed, but from the light which liberated them. This is in accord with the quantum theory of radiation. When a luminous quantum $h\nu$ (Article 537) strikes the surface, if an electron is ejected, its kinetic energy is given by Einstein's equation

$$\frac{1}{2}mv^2 = h\nu - w, \quad (1)$$

where w is the energy required to get the electron out of the metallic surface and is called the **work function**. Thus w acts as a tax on the incident quantum, and the kinetic energy of the ejected electron is less than that of the quantum which expelled it.

It is found by experiment that a positive charge on the photosensitive plate tends to diminish the effect, and if large enough, stops it altogether. This occurs automatically if the plate is insulated, for the loss of electrons gives it a positive charge that soon becomes large enough to hold them back by electrostatic attraction. The stoppage is not abrupt, however, showing that the escaping electrons have different velocities. The maximum velocity is attained by electrons whose w is very small, and then $\frac{1}{2}mv^2 = h\nu$, very nearly. If it takes V volts to stop these fastest electrons, their kinetic energy may be

measured by eV electron volts (Article 755), and if we set $w = 0$, equation (1) becomes

$$h\nu = eV \times 10^8, \quad (2)$$

when e is expressed in e.m.u.

768. The photoelectric cell. This device is much used today both for detecting and measuring either visible or ultraviolet light, and in such practical applications as television and the production of sounds from the film used in "talking" motion pictures. The discovery of a variety of compounds such as potassium hydride, far more sensitive to light (especially in the visible range) than pure metals, has made possible this remarkable development of the photoelectric cell. It is wonderfully adapted to its purpose, being extremely sensitive to minute changes of illumination. The current developed by a battery in series with such a cell varies directly as the intensity of the beam. Thus it reproduces electrically the variations of the light which falls upon it. Moreover, as it is almost instantaneous in its response (see Article 767), very rapid changes of illumination are accurately followed, including those due to the higher harmonics of sound waves when photographed on a film. In the more familiar types of cell, the light-sensitive material is coated over the inside of the tube, leaving a window of uncoated glass through which the light can pass. This layer forms the cathode, while a ring of some insensitive metal like copper forms the anode and is maintained at a positive potential with respect to the cathode.

A modern highly sensitive photoelectric cell made by the Bell Telephone Company† is fifty times as sensitive as the older potassium hydride cells. The anode A is a nickel wire in the axis of the cathode, as illustrated in Fig. 23. The cathode, which emits electrons when illuminated, is a sheet of pure silver curved into a half cylinder. Upon this base is formed a matrix of cesium oxide, silver oxide, and finely divided silver. Finally, after the tube has been exhausted, the matrix is covered with an adsorbed layer of cesium about one atom thick, using a technique too complicated for discussion here. Finally, if the tube is destined to carry relatively large currents, it is filled with an inert gas, like argon, at low pressure, so that positive ions may be produced by collision, as in the tungar rectifier.

Still another type of photoelectric cell depends upon a principle first observed by Grondahl and Geiger, American physicists. This



Fig. 23.

† M. J. Kelly, *Bell Telephone Laboratories Record*, October, 1933.

was later developed, by Schottky and Lange, in Germany, into the modern photronic cell, much used in photography for measuring the illumination of the "subject." It operates without an exhausted tube or a battery. The active material is a thin layer *B* of cuprous oxide (Cu_2O), formed on a base *A* of copper, as shown in Fig. 24. Over the oxide is deposited an excessively thin and fairly transparent layer *C* of

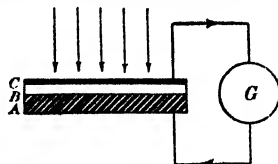
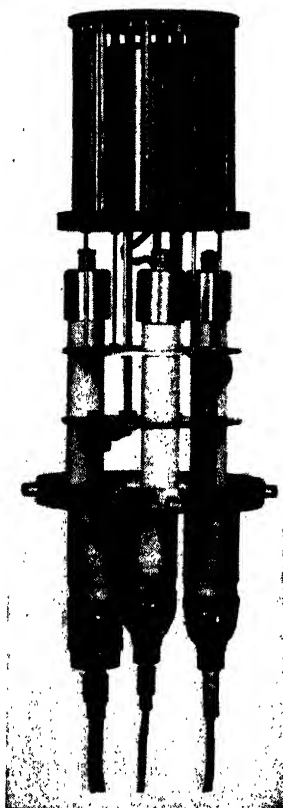


Fig. 24.

some metal. When illuminated, as indicated by the vertical arrows, an e.m.f., caused by the diffusion of electrons liberated in the photo-sensitive cuprous oxide, is set up across the boundary between *B* and *C*. Wires connected to the metal film and the copper plate complete the circuit, and a current flows externally from the surface film to the base plate.

A curious feature of the photronic cell is that in the dark it acts as a valve to currents produced externally. It stops a current flowing as indicated in Fig. 24, but allows it to flow in the opposite direction. In fact it was this valve action which Grondahl first developed to a high degree of perfection before the photoelectric e.m.f. was noticed. When this cell is used as a valve or rectifier, there is no upper layer *C*, and the valve action occurs at the layer separating the cuprous oxide and the metallic copper. In this layer a so-called "back wall" photoelectric e.m.f. is developed, but it is not nearly so strong as that developed between *B* and *C* in the "front wall" cell of Schottky and Lange, described above.



Courtesy General Electric Co.

Plate 25.

Cathode and grid structure of
G. E. thyatron.

769. The thyatron. Even gas-filled photoelectric cells can handle currents of but a few milliamperes at most, and usually operate at much lower values. In order to have light control

of lamps, switches, and so on, which need large currents, a relay called **thyatron** has been developed by the General Electric Company. It is a three-element tube filled with mercury vapor, through which flows a current of many amperes when the cathode is sufficiently heated and the grid is at a suitable potential.

If the grid potential is strongly negative, no current flows, but at a certain critical potential, usually a little below zero, the current starts and almost instantly reaches the full value corresponding to the temperature of the cathode, the anode potential, and resistance of the circuit. After the current is once established, the grid ceases to function as a control, because it becomes surrounded by a layer of positive mercury ions. However, if the applied potential is alternating, and if the grid potential is gradually lowered beyond the critical value, the current abruptly ceases when the applied e.m.f. passes through zero. In this way the thyatron acts as a faucet and turns on or stops a stream of electricity, instead of water.

The diagram in Fig. 25 shows how a thyatron T is used in connection with a photoelectric cell P . The function of the usual "A" battery is supplied by the A coils of the transformer M . The B coils act as the usual "B" battery to send a pulsating unidirectional current I through the thyatron and load L in the usual direction, as indicated by the arrow. When no light shines upon the cathode e of the cell, the "C" battery maintains the grid at a sufficiently negative value to prevent any discharge through T . But when e is illuminated, the

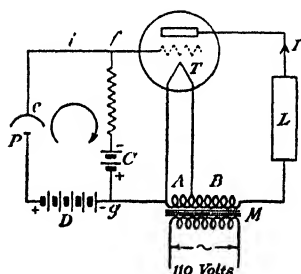


Fig. 25.

"D" battery sets up a flow i around the circuit $efgh$, in the sense indicated by the curved arrow. As the "D" battery has a higher voltage than the "C" battery, the drop across the resistance between f and C is reversed, giving f a higher potential than g , instead of a lower potential as is the case when C alone is acting. If the potential of f and the grid are raised above the critical value, the pulsating current I begins to flow and continues as long as e is illuminated enough to maintain the grid above that potential. Thus, by illuminating or interrupting the illumination of e , we control the photoelectric current i . This turns on or off the main current I , which may operate a thousand-watt lamp or a switch or other load.

SUPPLEMENTARY READING

- R. A. Hudson, *Electronics*, Wiley, 1932.
C. R. Underhill, *Electrons at Work*, McGraw-Hill, 1933.
R. A. Millikan, *Electrons (+ and -)*, University of Chicago Press, 1935.
J. H. Morecroft, *Electron Tubes and Their Applications*, Wiley, 1935.
Karl K. Darrow, *Introduction to Contemporary Physics* (Chap. 7), Van Nostrand, 1926.
J. A. Crowther, *Ions, Electrons and Ionizing Radiations*, Edwin Arnold, London, 1934.
K. Henney, *Principles of Radio*, Wiley, 1929.
Zworykin and Wilson, *Photocells and Their Application*, Wiley, 1930.
Hughes and DuBridge, *Photoelectric Phenomena*, McGraw-Hill, 1932.

PROBLEM

- *1. Calculate the wave length of a beam of light which liberates electrons from a zinc plate, if it takes 3 volts to stop the fastest. *Ans.* 4137 Å.

CHAPTER 57

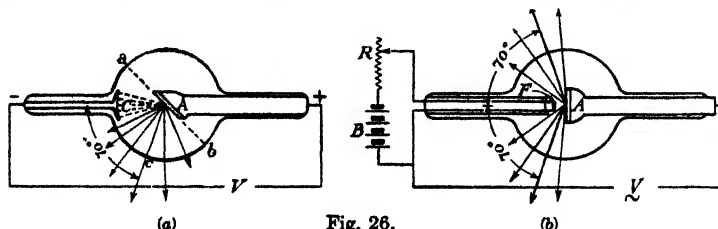
X-rays and Related Phenomena

770. Röntgen rays. Professor Röntgen of Munich, in 1895, discovered that a photographic plate in its holder was fogged when in the neighborhood of a Crookes tube, as a result of the discharge. This was later shown to be associated with the impact of the cathode rays upon the walls of the tube. If, however, some refractory metal of high atomic weight were used as a target, upon which the cathode rays might impinge instead of on the glass, the result was found to be much more powerful, and all such tubes are now constructed in this way.

The rays thus produced, unlike cathode and canal rays, are not corpuscular, and are not deviable in electrostatic or electromagnetic fields. They differ from light in being able to penetrate opaque objects, and are not readily stopped, or rather absorbed, except by matter of considerable density and thickness, such as sheets of metal, especially lead. The name X-rays was soon given to them because of their then unknown nature, and this term is still used, though their mystery has been largely explained.

In addition to their penetration, X-rays possess the power of ionizing gases to a high degree. They also produce chemical changes in certain substances like the emulsion of the photographic plate, and have profound physiological effects on animal and vegetable tissues, which may be either harmful, or, in treating certain diseases, highly beneficial.

771. X-ray tubes. It is an experimental fact, partly accounted for by theory, that X-rays are weakest in a direction opposite to the

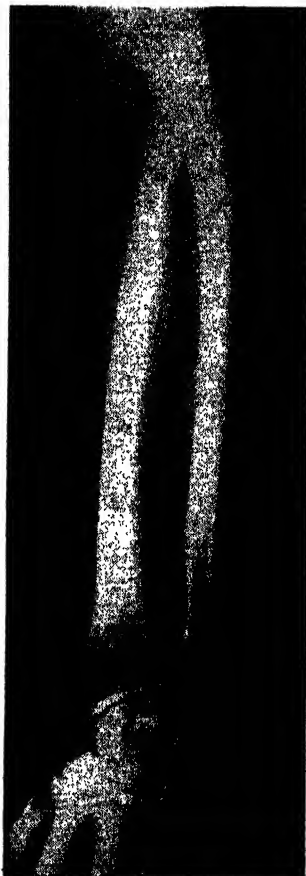


cathode stream. There is another minimum when rays are formed in the same direction as the cathode stream by passing through a

thin target. The maximum of intensity occurs at varying angles, depending on the voltage used and the material of the target, a fair average being about 70° from the cathode stream. Consequently, all the earlier X-ray tubes were designed with a target, inclined as in Fig. 26 (a). This results in a broader X-ray beam than if the target were normal to the cathode rays, but does not affect the direction of maximum intensity, indicated by the arrow *Ac*. The hemisphere of the bulb *acb* fluoresces under the bombardment of secondary electrons liberated from the anode by the impact of the primary cathode rays, indicated by the converging dotted lines.

In Fig. 26 (b) is shown a more modern type of tube known as the Coolidge tube, after its inventor, W. D. Coolidge, of the General Electric Company. Here the cathode consists of a coiled filament *F* of fine tungsten wire at the center of a concave cup. This filament may be heated to any desired temperature by the battery *B*, controlled by the rheostat *R*. The anode is faced with tungsten or molybdenum mounted on a massive rod which may be cooled by an internal stream of water, or air-cooled by vanes on the portion projecting outside of the tube. It comes quite close to the cathode, and the electron stream strikes it at right angles so that the X-rays emerge in a zone around the tube, as indicated by the arrows.

Coolidge tubes are so highly exhausted that no current flows until the filament is heated to redness, and then the intensity of the electronic current and resulting X-ray beam is adjusted as desired. The penetration or hardness of the rays is controlled by the impressed e.m.f., which determines the speed of the electrons. It may be either alternating or direct, as the tube has a valve action in common with other heated-cathode devices. When



Courtesy Department of Radiology,
Hartford Hospital.

Plate 26.

X-ray photograph of part of child's hand and arm, showing tissues and bone structure with unusual clearness.

used to examine tissues, bones, and so on, a minimum of about 35,000 volts is needed to penetrate the flesh. Bones are best seen when 45,000 volts are applied. Above 50,000 volts, the rays pass through the bones increasingly, and afford less and less contrast to the transparent tissues. The electronic current varies from a few milliamperes up to one ampere in exceptional cases with specially designed tubes.

The examination of bones and tissues just discussed is effected either by the use of a fluorescent screen, as described in Article 543, or by letting the rays fall upon a photographic plate, which is thus "exposed," and then developed in the usual manner.

772. The nature of X-rays. When these rays were first investigated, they were supposed to be a sharp "pulse" in the "ether" without any vibratory characteristics, somewhat like the sound produced by cracking a whip. This would be the logical consequence of the sudden stopping of the cathode-ray electrons when they strike the target. It was found that mirrors, prisms, gratings, and so forth, with which ordinary light is reflected, refracted, and diffracted, failed to produce the same effect on X-rays. This was supposed to confirm the "pulse" theory, for a disturbance without periodic vibrations would not behave like light, and would moreover have much higher penetration. But we now know that X-rays have a definite wave length and that their frequency of vibration depends upon the velocity of the moving electrons whose impact with the target produces them, with certain limitations imposed by the atoms of which the target is made. They are electromagnetic vibrations like light, though their wave lengths average from five to ten thousand times shorter.

773. X-ray spectra. The rays emitted from the target of an X-ray tube form a continuous spectrum like that of light from an incandescent solid. They also exhibit "lines" characteristic of the metal of which the target is made.

The continuous spectrum has a very sharp upper limit which depends upon the voltage applied to the tube. The potential difference gives the electron an energy of eV electron volts, which is just capable of liberating a quantum $h\nu_m$ from the target, where ν_m is the maximum frequency due to the fastest electrons in the cathode beam. This is the inverse of the photoelectric effect, as expressed by equation (2) of Article 767, for in both cases the work function w vanishes for the highest frequency. The inverse relation, $eV \times 10^8 = h\nu_m$, was discovered by Duane and Hunt in 1915, and is often referred to as the Duane-Hunt relation. As both the potential and the limiting frequency ν_m may be accurately measured, this relation enabled its authors to

determine Planck's constant h with great precision. Their result agrees closely with the accepted value obtained by Millikan in another way.

Below the maximum frequency of the upper limit, the target emits X-rays of all frequencies, extending the spectrum indefinitely into the region of longer wave lengths. With the usual thick targets, the *intensity* of the beam is greatest at frequencies somewhat lower than the maximum, and then falls off gradually toward the long-wave end of the spectrum.

Characteristic X-rays depend upon the metal of the target, and are imposed, as it were, upon a background of the continuous spectrum. To excite these "lines," the electron must have sufficient kinetic energy to supply the quantum equivalent to that line of a series which has the highest frequency. When this is reached by raising the voltage, then all the lines of the series appear at once. With reference to a particular series of lines observed in the characteristic spectra of different elements, it is found that increasing voltages must be used to excite this series in elements of increasing atomic number. This means that the bombarding electrons have to travel faster to be able to excite the characteristic lines of heavier atoms. The quanta emitted are thus larger, and the resulting frequencies are higher for corresponding lines as we ascend the scale of the atoms.

A simple calculation shows that it takes 12,360 volts acting upon an electron to produce a wave length of one angstrom (10^{-8} cm), but as many characteristic waves emitted by the heavier atoms are even shorter than this, much higher voltages are needed to give the electron enough energy to excite the atom sufficiently to enable it to radiate these short wave lengths.

774. Secondary X-rays and corpuscular radiations. When X-rays fall upon a "radiator," they give rise to two types of secondary rays. One of these types consists of *fluorescent* X-rays, so called because, as with ordinary fluorescence, the frequency of the excited rays depends upon the nature of the radiator, and is always lower than that of the incident beam, in accordance with Stokes' law. In fact, they are nothing but an example of fluorescence in a realm of very short wave lengths, and the incident quantum must have more energy (that is, higher frequency) than that of the fluorescent rays to be generated. The other type of secondary rays is said to be *scattered*. These have the same frequency as the original rays, and are produced somewhat as visible light is scattered by a rough surface.

Both fluorescent and scattered X-rays are associated with the emis-

sion of electrons from the radiator. In the case of fluorescent X-rays, the electrons are emitted as in the photoelectric effect, and Einstein's equation is applicable. The fastest electrons have the same kinetic energy as those in the primary beam, given by

$$\frac{1}{2}mv^2 = h\nu = eV \times 10^8,$$

where V is the voltage impressed upon the X-ray tube. This is really the same equation as that which governs the maximum energy of electrons emitted in the photoelectric effect with visible light. Here, however, the size of the quantum is so great that the small loss of energy w needed to get the electron out is negligible in comparison to $h\nu$.

775. The Compton effect. In 1922, Professor A. H. Compton, of the University of Chicago, discovered that scattered X-rays, in addition to exhibiting the frequency of the original beam, were accompanied by a ray of slightly lower frequency. This phenomenon, known as the *Compton effect*, has a certain similarity to fluorescence, but it is really very different, because the frequency of the scattered rays observed by Compton, and also by Raman (Article 546), has nothing to do with the substance of the secondary radiator.

The lowering of the frequency observed by Compton means that the quanta of the scattered X-rays are always smaller than the primary quanta which produced them. That is, $h\nu - h\nu' = \epsilon$, where ν' is the frequency of the secondary rays, and ϵ is a small amount of energy taken up by an electron which recoils as a result of the encounter with the quantum. This equation expresses the law of the conservation of energy applied to quanta, and the principle involved has done much to broaden the applications of the quantum theory.

776. Crystalline structure. The discovery that X-rays have definite wave lengths, and so behave like light, is due to a suggestion made in 1912 by von Laue, of Munich, that crystals, because of their very fine and regular structure might act toward X-rays as gratings do toward visible light. This experiment was carried out by Friedrich and Knipping, who caused a fine pencil of X-rays to pass through a crystal and then fall on a photographic plate. The result entirely justified von Laue's prediction, for instead of a single spot on the plate, there

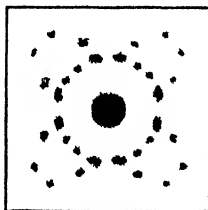
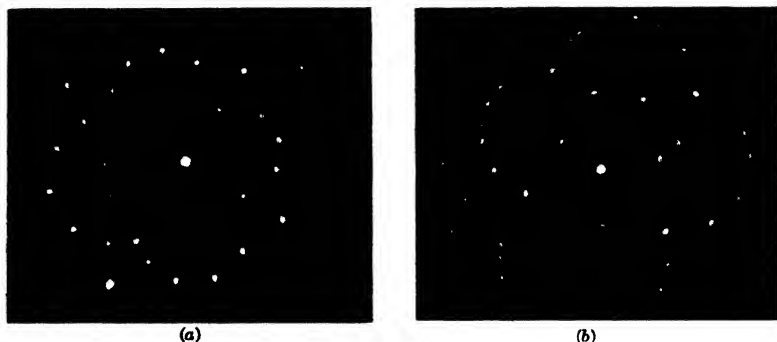


Fig 27.

was a series of spots arranged in a geometrical pattern, as in Fig. 27.

This discovery was immediately followed up by W. H. Bragg and

his son, W. L. Bragg, who made use of *reflection* from crystal surfaces, as suggested by the younger Bragg, instead of *transmission* through the crystal. The Braggs' method proved a much more effective way of diffracting X-rays, and led to their brilliant researches on the struc-



Courtesy Professor Arthur Wadlund, Trinity College.

Plate 27.

"Laue photographs." (a) X-ray diffraction pattern taken through a crystal of rock salt (NaCl), normal to cleavage plane. (b) X-ray pattern through a crystal of calcite (Iceland spar, CaCO_3), normal to cleavage plane. In both (a) and (b) the central spot is due to the undiffracted pencil of X-rays.

ture of crystals and other substances. In this way they demonstrated the regular grouping of the atoms in crystalline bodies, and that they are built up on what is known as a *space lattice*.

A simple cubical lattice is one made of elementary cubes having an atom at every corner, as in Fig. 28 (a). A face-centered cubical lattice (b) has, in addition, an atom at the center of each face, while a body-centered lattice (c) has an atom at the center of each elementary cube. It will be seen that the latter system has just twice as many atoms as (a), because if we consider a lattice composed of many cubes, each atom in (a) takes part in eight cubes of which it is a corner; therefore there are as many cubes as atoms. In (c), however, we have added one atom to each cube, thus doubling the total number and forming a new lattice that interpenetrates the original one. The system (b) has added six atoms to the cube shown, but each belongs to

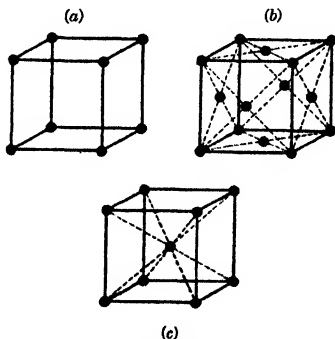


Fig. 28.

another cube as well; therefore, really only three have been added to the original single atom per cube, and the total number per cube has been quadrupled.

One of the simplest of such lattices is the crystal of rock salt, NaCl , whose elementary cube is made up of alternate atoms of the two

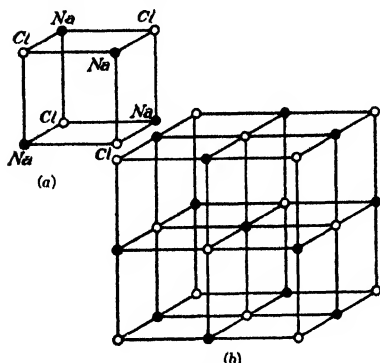


Fig. 29.

elements, as shown in Fig. 29(a). But if eight such cubes are taken, we find a face-centered cubical lattice marked by the chlorine atoms, and a similar interpenetrating lattice of sodium. These lattices have elementary cubes eight times as large as in (a), but being face centered, they contain four times as many atoms of either element. This is the same as saying that the elementary cube in (a) contains half an atom of either kind.

The lattice structure shows that a crystal is made up of a great variety of parallel planes, horizontal, vertical, and diagonal, as defined by the atoms, and they are therefore suitable for causing diffraction in three dimensions, rather than in a single plane as with the ordinary ruled grating.

777. Crystal gratings. The explanation of the process of producing the spectrum of a source of X-rays differs from the theory of the reflection grating with ordinary light. Unlike light, the X-rays penetrate the crystalline grating to a depth of thousands of planes, and what is equivalent to reflection occurs only at the points occupied by the atoms which first absorb and then re-emit the energy that falls on them. In order to derive the equation of such a grating, it is necessary to consider only the two planes shown in Fig. 30. The difference in path between the two rays AA' and BB' is clearly equal to pnq , where p and q are obtained by dropping perpendiculars from m on Bn and $B'n$ respectively, so that down to mp and after mq , the two paths are the same in length. But the angles nmp and nmq , are equal to θ having their sides mutually perpendicular. Therefore $np =$

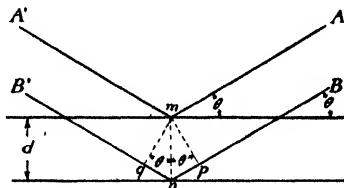


Fig. 30.

$nq = d \sin \theta$, where d is the distance between planes. Then the difference of path δ is

$$pn + nq = 2d \sin \theta. \quad (1)$$

The two rays indicated in the diagram will be able to interfere after this "reflection" if δ involves an odd number of half wave lengths, and will reinforce each other if it contains an even number. As in the ordinary grating, a few thousand such planes will prevent all but a very small band of wave lengths from being reflected at a given angle, and the relation which determines this selective reinforcement is

$$2d \sin \theta = 2n \frac{\lambda}{2} = n\lambda. \quad (2)$$

In order to use this equation for measuring wave lengths, it is necessary to know d . This quantity may be calculated for rock salt as follows: Since each elementary cube contains half an atom of each element, or half a molecule of NaCl, the mass of an elementary cube is $M/2N$, where M is the molecular weight of sodium chloride and N is Avogadro's number, or the number of molecules in the mass of a substance numerically equal to its molecular weight (gram molecule). But the atomic weights of sodium and chlorine are 23 and 35.5 respectively; therefore M , their sum, is 58.5, and $N = 6.06 \times 10^{23}$, so the cube's mass is known. The volume of the cube is obtained by dividing its mass by the density D of rock salt, which is 2.17. Therefore

$$d^3 = \frac{M}{2DN}, \quad (3)$$

where d is the edge of a cube and the distance between the planes we are considering. The calculation thus indicated gives $d = 2.814 \times 10^{-8}$ cm, a constant that is very important in X-ray spectroscopy.

778. Meaning of the spectra. The spectra obtained by the method just outlined are important from two points of view. They serve to analyze X-rays, with a crystal of known constant d , or they may be used as a basis for the study of crystal structure, when X-rays are employed whose wave length has previously been measured. The analysis of X-rays has been of the greatest value in getting at the structure of the atom, for the rays characteristic of the metal of the target are produced deep down within the atom, instead of at its surface, as is usually the case with the emission of visible spectra. The use of X-rays in the study of crystal structure has resulted in a re-

markable analysis of the grouping of the atoms not only in crystals, but also in such substances as oils and other hydrocarbons, and has accounted for some of their physical properties, as well as vindicating the more or less hypothetical molecular structures assumed for them by the chemists.

779. The Wilson cloud chamber. The method used by J. J. Thomson in 1899 to measure the charge of the electron was based on a discovery, made by C. T. R. Wilson two years earlier, that both positive and negative gas ions may act as nuclei of condensation for water vapor. The ions were produced by X-rays in a chamber containing water vapor, and the volume of the gas was suddenly expanded by 25 per cent. This expansion cooled the gas, but was too small to produce condensation in dust-free air unless ions were present. If they were, water vapor formed around them, and a cloud appeared that slowly settled to the bottom of the vessel. From its rate of descent and other considerations, the charge on each droplet could be calculated. Although the cloud chamber was superseded as a means for obtaining the value of e , it is still of great practical value in the observation of ionization by X-rays, fast electrons, and fast helium ions (alpha rays). A simplified form of the apparatus is shown in Fig. 31 (a). The space S is saturated with vapor from the water supply W , whose level is raised to the dotted line by squeezing the bulb B . The bulb is then released, and the necessary sudden expansion occurs.

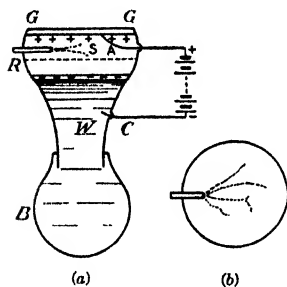


Fig. 31.

The alpha rays are ionized helium atoms that plow short but destructive trails through the air and form ions from its molecules at very short intervals. These ions in turn serve as condensation nuclei upon which water vapor condenses at the moment of expansion, and the tracks appear as fine, luminous, nearly straight lines composed of tiny beads closely strung together, as indicated in (b), which is the ionization chamber as seen from above.

An important accessory to this apparatus is the battery of about

90 volts whose terminals are connected to the moist lower surface of the glass *GG* by the wire *A*, and to the water by another wire *C*. Thus, the charges on the lower surface of the glass and the upper surface of the water produce a field of force which sweeps out all ions from the space *S* except those that happen to be formed at the moment of expansion.

The positive α rays almost always ionize an atom by passing through its outer structure without being appreciably deviated. But occasionally they make a head-on collision with the nucleus, and the sharp deflection which then occurs is indicated by the bend in the track shown at *p* in Fig. 32 (*a*). In general, the atom collided with acquires sufficient momentum to ionize on its own account. It then forms a track of its own which is occasionally sufficiently distinct to be seen as a spur, such as the one indicated at

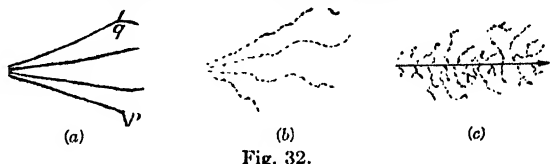


Fig. 32.

q. Fast-moving electrons produce tracks similar to α tracks, but with the visible droplets more widely spaced because they are less effective as ionizing agents. These tracks are also more irregular and longer, as seen in (*b*), because the electron is more easily deviated in passing through the atom it ionizes, though it has greater penetration. The path of X-rays is shown in (*c*). These rays, indicated by the arrow, ionize the gas for the most part indirectly. That is, they produce a relatively small number of positive and negative ions directly, and then these rather slowly moving particles produce many more ions by collision along very irregular paths, as indicated.

780. The positron. In September, 1932, C. D. Anderson, of the California Institute of Technology, reported the discovery of positive electrons. These particles were first discovered in connection with observations on cosmic rays by means of the Wilson vapor tracks. The first, now historic, photograph of the path of a **positron**, as positive electrons are commonly called, was obtained on August 2, 1932, and is sketched in Fig. 33 (*a*). The cosmic rays descending vertically appear to have produced a positron from an atom of the gas in the chamber at *P*, and endowed it with enormous energy. Its initial velocity was nearly normal to the lead screen *LL*, but a transverse magnetic field of 15,000 oersteds caused it to curve toward the left. It then encountered the screen, which was 6 mm thick. This slowed it down, as is shown by the increased curvature of the track beyond

the screen, proving that it must have been moving from P toward LL . The direction of its curvature in the field directed away from the observer proves it to have had a positive charge. Furthermore,

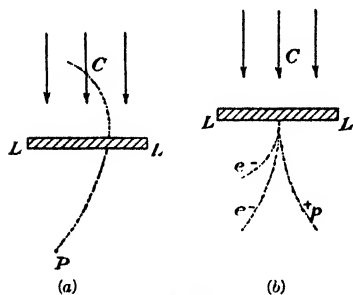


Fig. 33.

the length and general appearance of the track rule out the possibility of a moving proton, because the track of a proton would be much shorter and the droplets composing it would be much closer together.

Since Anderson's first photograph was taken, a very great number have been obtained, notably at Pasadena and the Cavendish Laboratory

of Cambridge University. In some cases, both electrons and positrons are produced simultaneously by the cosmic rays, as indicated in Fig. 33 (b). In this case the source is within the lead sheet L subjected to the cosmic rays C . The track p is that of a positron curved by a field directed away from the observer, and the tracks e are those of electrons. The electron tracks are more curved than the positron track, showing that the positron has a higher speed. This is to be expected, because an electron leaving the positively charged nucleus would be attracted back, while a positron would be repelled.

The curvature of the positron tracks in a field of known strength enables us to calculate the velocity of the particle and, therefore, its energy in electron volts. The energy of the positron of Fig. 33 (a), assuming its charge to be the same as that of an electron, was calculated in this way, and found to be 63 million electron volts before passing through the lead plate, and 23 million afterward.

Positrons are also produced simultaneously with electrons by bombarding screens of various materials; such as cellophane, aluminum, and lead, with gamma rays from radioactive substances (especially thorium B and C), and with neutrons produced in a manner described in Article 802. Gamma rays (Article 812) are intermediate between cosmic and X-rays as regards penetration. The electron-positron pairs are produced by gamma rays having an equivalent energy of a million electron volts or more, and the rate of their production from a given source varies as the square of the atomic number of the bombarded substance. Occasional "bursts" or "showers" of electrons are also produced by gamma rays and by neutrons. These indicate a collision of explosive violence.

781. Cosmic rays. It has long been known that a charged electroscope discharges even when the greatest care is taken to insulate the leaves. The discharge is therefore due to ions in the atmosphere. In 1903, Rutherford and McLennan found that when an electroscope is surrounded by a heavy metallic shield, the leak still persisted, though it was not so rapid as before. This seemed to indicate that some of the ions must be due to a form of radiation even more penetrating than X-rays or the gamma rays of radium. These new rays were further investigated by Gockel, and later by Hess and Kohlhorster (German physicists), who carried an electroscope up in a balloon and found that at great altitudes the intensity of the rays steadily increased. This proved that the rays came from outer space, having been partially absorbed in the upper air. Thus they came to be called **cosmic rays**.

Then Millikan attacked the problem. He demonstrated the amazing penetration of these rays, and found that they could penetrate several feet of lead and discharge an electroscope sunk deep in the water of a lake. He also showed that they do not appear to come from the sun or milky way, but apparently fall upon the earth from every portion of the heavens.

At first Millikan believed that cosmic rays were electromagnetic vibrations like light, but of much higher frequency. He called them the "birth cries" of new-born atoms formed in interstellar space. According to the quantum theory, quanta of very great energy and high frequency would be released with loss of mass when protons, neutrons, and electrons unite to form an atom. This high frequency of the resulting photon would give it high penetration, as we have seen. But calculation did not yield a value high enough to account for the observed penetration. Moreover, A. H. Compton, who had begun serious investigations about 1928, found that the intensity of cosmic rays varied with the latitude, being stronger nearer the poles. This would not be true of electromagnetic waves, but would be true of charged particles such as high-speed electrons, whose penetration could be explained if they had a velocity nearly equal to that of light.

At present there is a general agreement that whatever the cosmic rays may be in outer space, here where we can observe them they are charged particles, probably accompanied by electromagnetic waves. The electromagnetic waves may have produced the corpuscular rays by ionization in the upper strata of the atmosphere, or they may be the result of collisions of purely corpuscular cosmic rays with atmospheric molecules. The latter seems the more probable hypothesis,

but it comes no nearer to accounting for the origin of these tremendously energetic radiations.

Cosmic rays are important both because they may help us explain the evolution of the cosmos, and because it seems quite likely that they have played an important role in the evolution of organic life upon the earth.

SUPPLEMENTARY READING

- B. L. Worsnop, *X-rays*, Methuen, London, 1930.
Karl K. Darrow, *Introduction to Contemporary Physics* (Chapters 8, 9, 10), D. Van Nostrand, 1926.
C. F. Meyer, *Diffraction of Light, X-rays and Material Particles*, University of Chicago Press, 1934.
G. W. C. Kaye, *X rays*, Longmans, Green, 1929.
—, *High Vacua*, Longmans, Green, 1927.
J. T. Randall, *The Diffraction of X-rays and Electrons*, Wiley, 1934.
Harvey B. Lemon, *Cosmic Rays Thus Far*, W. W. Norton, 1936.

PROBLEMS

1. Calculate the angle θ (Fig. 30) of the first-order X-ray spectrum from rock salt, of the K series α_2 line of calcium. ($\lambda = 3.359 \text{ \AA}$) *Ans.* $36^\circ 38' 38''$.
2. What is the wave length of an X-ray first-order spectral line reflected from rock salt at an angle of $2^\circ 10' 23''$? *Ans.* 0.21341 \AA (α_2 line of K series of tungsten).

CHAPTER 58

Atomic Structure

782. Periodic series of the elements. It has long been known that the elements of which all matter is composed follow each other in a natural series of ascending atomic weights; and, as was shown by Mendeléef in 1870, certain chemical properties repeat themselves after advancing eight steps in the series. Thus the chemically similar elements, lithium, sodium, and potassium, are numbers 3, 11, and 19 in the series, differing by eight in their serial order. The same is true of the gases, helium, neon, and argon, whose serial numbers are 2, 10, and 18. These groups not only have similar chemical properties, but their members resemble each other in their physical behavior, such as their boiling and freezing points, atomic volume, and various mechanical and optical properties.

It was also noticed that many atomic weights became exact integers, if oxygen were taken arbitrarily as 16, and others were more nearly so than the laws of probability would have predicted. Moreover, the atomic weights of all the lighter elements were seen to be approximately twice the value of their serial numbers.

Following out these indications, Mendeléef was able to predict the existence of the elements gallium, scandium, and germanium, not then known, because they were needed to fill certain gaps in his series, and he correctly predicted even their chemical properties.

Such evidences of regularity and system long ago suggested the possibility that the elements themselves were made up of some still more elementary substance, in an arrangement of increasing complexity for the heavier atoms. Thus the name *atom*, meaning *indivisible*, would no longer be strictly applicable.

The discovery of radium and other radioactive substances, which exhibit a process of transmutation from element to element, gave this idea a great impetus, and atoms are now known to be made up of at least three, or even more, fundamental entities (protons, electrons, and neutrons) in varying quantities and arrangements. In spite of this fact, the name *atom* may still be used, because, though not indivisible, it is the smallest particle into which an element can be divided and still retain its identity.

783. Atomic number. Although there are various ways of picturing the atom, it is now generally agreed that it consists of a very small positive nucleus surrounded by negative electrons, and that their number determines its position in the atomic series. Thus hydrogen, number 1, has one electron outside the nucleus; oxygen, number 8, has eight exterior electrons, and uranium, the heaviest atom, number 92, has ninety-two exterior electrons. There is further evidence, as we shall see, which leads us to regard the serial, or **atomic number** (denoted by Z), as of greater significance than the atomic weight in the determination of an element's properties. In four cases, two adjacent elements in Mendeléef's series have been reversed by evidence derived from their spectra, such as potassium and argon, whose atomic weights are 39.1 and 39.88, respectively, but whose X-ray spectra show that the former has nineteen outer electrons and the latter only eighteen. Thus their atomic numbers, 19 and 18, reverse their earlier sequence, which placed argon higher in the scale.

784. Mass of the atom. Even in the case of the heaviest atom, uranium, with ninety-two outer electrons, their total mass is only $92/1835$ of the mass of the lightest atom, hydrogen. As the atomic weight of uranium is 238.2, it is clear that the electrons form a very minute portion of its mass. It follows that the mass of the atom is concentrated mainly in its nucleus, which is extremely small, and has a density thousands of times greater than that of any known material. This mass advances from atom to atom by increments sufficiently regular to suggest as a first approximation that it might be made up of hydrogen nuclei, each of an atomic weight not far from unity.

785. The proton. The nucleus of hydrogen is called a **proton**, and has a positive charge equal to that of the electron which is associated with it in the atom. It is, therefore, the same as the hydrogen ion, because any atom becomes ionized by losing an electron, and as hydrogen has but one to lose, the result is the electropositive proton. This is regarded as the ultimate stable particle of positive electricity, just as the electron is of negative electricity, but its mass is 1835 times as great, constituting practically all of the atom's mass. Deducting the electron's "atomic" weight, 0.00055, from the atomic weight of hydrogen, 1.0081, we obtain 1.0076 as the "atomic" weight of the proton.

786. The neutron. In 1932, Chadwick, of Cambridge University, concluded that certain very penetrating rays first observed by Bothe and Becker in 1930 were not electromagnetic vibrations like X-rays, as had been supposed, but were composed of uncharged particles of

about the same mass as a proton. The methods by which neutron rays are obtained and the experimental evidence for their supposed nature will be discussed farther on. For the present it is sufficient to state that their atomic weight is about 1.0091, which is slightly larger than that of the proton calculated above.

787. The nucleus. The next heavier atom after hydrogen is that of the gas deuterium. It is a newly discovered form (isotope) of hydrogen, and is described by the symbol ${}_1\text{H}^2$. This notation indicates an atomic weight of 2 (approximately), and an atomic number of 1. Thus it has one outer electron and therefore has the same chemical properties as hydrogen. Deuterium combines with oxygen to form water of density 1.1. Its nucleus, called a **deuteron**, is a union of a proton with a neutron. An atom ${}_1\text{H}^3$ also exists, but little is known about it as yet.

After the isotopes of hydrogen, we come to helium, with 2 as its atomic number. Its nucleus consists of two neutrons and two protons which hold two orbital electrons in equilibrium. The atomic weight of the helium nucleus is 4.0041 when oxygen is taken as 16, instead of $2 \times 1.0091 + 2 \times 1.0076 = 4.0334$. This discrepancy of 0.0293 is accounted for by a loss of mass in the process of formation, which must have been accompanied by a tremendous evolution of energy. Such a violent union of its constituent particles would account for the extraordinary stability of the helium nucleus.

The stability of any nucleus heavier than that of hydrogen is not easily accounted for, because the two or more protons it contains would be expected to fly apart under the mutual repulsion of like charges. Bieler and Chadwick, in Cambridge University, first showed that the inverse square law of electrostatic repulsion between protons, when they came very close together, was no longer valid. These protons seemed to behave somewhat like magnetic doublets whose attraction varies inversely as the fifth power of the distance. Then in May, 1936, three investigators of the Carnegie Institution in Washington—Tuve, Heydenburg, and Hafstad—announced the results of experiments indicating a force of attraction between protons, when close together, which is very much greater than the electrostatic repulsion between them at the same distance. But this new force falls to zero with the distance much more rapidly than it would in accordance with an inverse square law. Such a force is negligible except at very close range. But at distances of the order of 10^{-12} cm, it is forty times the electrostatic repulsion between two protons the same distance apart. These results were obtained by bombarding hydro-

gen atoms with very high speed protons and measuring the angles through which they were deflected when in collision with the hydrogen nuclei. From these angles, measured at varying speeds, the forces involved were calculated.

The helium nucleus is also known as a **helion** and as an **alpha particle**. Because of its stability it is regarded, after electrons, protons, and neutrons, as a fourth "building block" of matter. This view is justified by the spontaneous emission of alpha particles from the radioactive elements. It is also significant that the nuclei of elements like carbon and oxygen, whose atomic weight is evenly divisible by four, are peculiarly stable. This suggests that they are made up exclusively of alpha particles.

The next heaviest atom is lithium, of atomic number 3, and atomic weight 6.940. If we assume the same hypothesis as was used above with helium, the nucleus of lithium would contain one alpha particle, one proton, and two neutrons, which may be written $\alpha + p + 2n$. The helion gives it two positive charges, the proton one, and the two neutrons none. These three unit charges hold three orbital electrons in equilibrium, thus forming an atom whose atomic number is 3 and atomic weight about 7. In this way we can build up heavier and heavier nuclei by choosing the requisite number of helions, protons, and neutrons so as to give the proper atomic weight and a resultant positive charge totaling that of a number of electrons equal to the atomic number of the element.

788. The isotopes. From what has been said, it is evident that by a proper combination of protons, neutrons, and helium nuclei, atoms may be obtained having either different atomic numbers with the same atomic weight, or the same atomic numbers with different atomic weights. Both of these exist. The former will be explained in the next chapter. As an illustration of the latter, known as **isotopes**, consider the gas neon. Its atomic number is 10, which means that its nucleus must have ten elementary positive charges. Its atomic weight is 20.2, which would seem to indicate five helions, leaving ten unbalanced charges to neutralize those of its ten orbital electrons. But it is also possible to imagine the nucleus made up of five helium nuclei and two neutrons, which would mean an atomic weight of 22, and the same positive charge as before. These two forms of neon are known to exist, and they must both be present in the ordinary gas, of atomic weight 20.2, in proportions of about ten of the first kind to one of the second, so that the mixture may have the desired weight.

As we ascend the scale of atomic weights, the isotopes become more frequent. Chlorine has two isotopes, of atomic weights 35 and 37, giving when mixed the atomic weight of the gas as 35.4. Krypton has six isotopes lying between 78 and 86, which result in an atomic weight of 82.92. Lead (atomic weight 207.2) has six, lying between 206 and 218, and there are many more similar cases, all of which tend to account for the deviation of atomic weights from whole numbers.

789. Mass spectra. The chemical behavior of an element depends almost entirely on its outer electrons; therefore isotopes cannot be separated or identified by chemical processes, except with great difficulty. In a few cases they have been separated by taking advantage of slight differences in their physical properties, such as the rates of diffusion of gases, which depend upon their atomic weights. But by far the best method is to form positive rays, similar to canal rays, of the element to be investigated.

The positive ions which pass from a heated anode coated with a metallic oxide constitute rays of positively charged atoms of the metal in question, and like canal rays, they may be deviated in electrostatic and electromagnetic fields. The amount of deviation depends upon their speed and mass, and a knowledge of the former gives us a means of calculating the latter.

These spectra were first studied by Sir J. J. Thomson, and later by F. W. Aston, both of Cambridge University. Aston's method consists in a very ingenious combination of crossed fields acting one after the other on the moving ions. This results in very beautiful "mass spectra" in which each "line" is a spot on the photographic plate where ions of only one atomic weight have struck. In this way, if an element has four isotopes, four distinct spots are obtained; and by measuring their position, the atomic weight of each isotope is found.

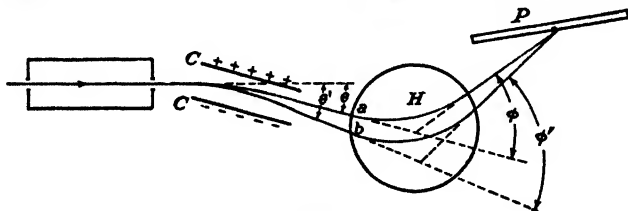


Fig. 34.

In Fig. 34, the electrostatic field is seen to be produced by the charged condenser CC , and the magnetic field H normal to the diagram is produced between magnetic poles indicated by the circle. A positive ion traveling between the charged plates is bent through

an angle θ , and after leaving the field, moves along the straight line a . A slower-moving particle of the same mass is bent through a larger angle θ' and follows the line b . On entering the magnetic field both particles are curved upward through angles ϕ and ϕ' , and with proper adjustment of the various distances and fields, they are made to strike the photographic plate P at the same point. Particles of a different mass are similarly concentrated, but at a different point.

790. Recent mass spectrographs. In the last few years a number of mass spectrographs different in principle from Aston's have been successfully used. Several of these have proved superior to the earlier type in certain particulars such as greater intensity, resolving power, or precision. One of these was devised by A. J. Dempster, of the University of Chicago. In his apparatus, now widely used, positive ions are first accelerated in an electrostatic field of V volts, thus giving all the particles the same kinetic energy, which may be calculated by

$$\frac{1}{2}mv^2 = eV, \quad (1)$$

where e is the ionic charge. These ions are then passed through a transverse magnetic field, somewhat as is shown in Fig. 7 (Article 752), but deviated through 180° . Under these conditions a sharp focus of all particles having the same momentum is obtained. Then, using equation (1) of Article 752, in conjunction with (1) above, we obtain

$$e/m = 2V/B^2r^2, \quad (2)$$

from which e/m may be found in terms of measurable quantities.

A mass spectrograph of remarkable resolving power and high precision was invented in 1933 by K. T. Bainbridge, now of Harvard University. The method used is to pass the positive ions through crossed fields similar to those shown in Fig. 8 (Article 753), with a slit at F which passes only those within a small range of velocities. After that they enter a magnetic field which sends them around a circular path until they strike the photographic plate. This, as in Dempster's method, focuses those of the same mass in a single line, and the dispersion is so great that differences of the order of the mass of an electron are clearly indicated. Thus comparison of the masses of isotopes is made with high precision.

791. Atomic models. After the discovery that atoms are made up of a positive nucleus surrounded by electrons, there were many attempts to picture their arrangement by means of some model making use of the laws of mechanics and electromagnetism. These were

unsuccessful because in the microcosm of the atom, the classic principles apparently are not valid, or must at least be modified under conditions which seem to be different from those which control large electric charges and masses involving great aggregates of electrons and protons.

The first atomic model satisfactory to physicists was proposed by Rutherford and later developed by Bohr. This atom immediately became popular because it went a long way toward accounting for a wide variety of observed facts. But even so, it is only a partial solution, and at present, owing to the discovery of the wave character of both electrons and protons, it seems necessary to abandon the attempt to form a definite picture of atomic structure. Its problems are best attacked by a combination of the quantum theory and a new branch of mathematics known as *wave mechanics*. This reduces the atom to a complex of waves not easily represented by a mechanical model, as was possible with the Bohr atom. But in order to understand the newer concepts, it is necessary first to have a picture of the atom from which these concepts were built up, especially as the picture is still valuable up to a certain point, and may be thought of as a "first approximation" toward solving a very elusive and subtle problem.

792. The Bohr atom. In 1913, the Danish physicist Niels Bohr proposed a hypothetical atom which now bears his name. According to his theory, later extended by Sommerfeld, the outer electrons move in orbits about the nucleus, like the planets around the sun. In order that this may be possible, however, Bohr was forced to make the assumption that although such a rotating planetary electron is being constantly accelerated toward the center, it does not radiate energy. This radiation should occur in accordance with classical electrodynamics whenever a charged body changes its velocity either in direction or in scalar magnitude. If the rotating electrons did radiate energy, the loss entailed would cause them to slow down and ultimately fall into the nucleus, thus resulting in the annihilation of the atom. Such a process would, moreover, yield a continuous spectrum instead of the observed bright line spectra characteristic of the various elements.

To account for the systematic arrangement of the lines of the spectra as observed in the Balmer series of hydrogen, and other similar series, Bohr made two other assumptions depending upon the quantum theory. These three postulates may be summarized as follows:

1. The orbital rotation *without radiation* follows the Newtonian laws, so that the force of attraction between the nucleus and the electron is equal to the centrifugal reaction of the latter, or

$$\frac{(Ze)e}{r^2} = m\omega^2 r, \quad (1)$$

where Z is the atomic number, Ze the nuclear charge, r the radius of the orbit, and $\omega = 2\pi f$, where f is the orbital frequency.

2. The number of orbits about a given nucleus is limited. In each possible orbit, the angular momentum of the electron is an integral multiple of $h/2\pi$, where h is Planck's constant. In a circular orbit whose radius is r , the angular momentum of a mass m is $m\omega r^2$; so Bohr's second postulate may be stated by

$$m\omega r^2 = nh/2\pi, \quad (2)$$

where n is any integer. This is called the **quantum condition**.

3. Radiation takes place only when an electron shifts from one orbit to another of smaller radius. If W_1 represents the total energy (kinetic and potential) of the electron in an inner orbit, and W_2 its energy in an outer orbit, then the energy released by the shift is $W_2 - W_1$. This energy is radiated as a quantum $h\nu$; therefore

$$W_2 - W_1 = h\nu. \quad (3)$$

In order to radiate, an atom must first absorb energy sufficient to lift an electron from an inner to an outer orbit, or level. This may be supplied by the absorption of a photon of energy $h\nu$, or by the impact of an electron whose energy eV exceeds the critical value, as pointed out in Article 773.

The total energy W of an electron in its orbit is calculated as follows: If we multiply equation (1) by r , and substitute $v = \omega r$, we obtain

$$W_K = \frac{mv^2}{2} = \frac{(Ze)e}{2r}, \quad (4)$$

where W_K is the kinetic energy. This shows us that the kinetic energy is greatest for orbits nearest the nucleus.

The potential energy, W_P , of two charges separated by a distance r , is equal to qq'/r , as was shown by equation (3), Article 587. In the case of a nucleus and electron, the charges are of opposite sign; therefore

$$W_P = -\frac{(Ze)e}{r}. \quad (5)$$

Then the total energy, $W_K + W_P$, is given by

$$W = \frac{(Ze)e}{2r} - \frac{(Ze)e}{r} = -\frac{(Ze)e}{2r}. \quad (6)$$

This relation tells us that the total energy of an orbital electron is *numerically* equal to its kinetic energy, but has a negative sign, because its negative potential energy is twice as great as its positive kinetic energy.

The radius of an orbit is found by eliminating ω between (1) and (2). The result is

$$r = \frac{n^2 h^2}{4\pi^2 (Ze)em}. \quad (7)$$

It follows that the radii of the orbits of a given atom are to each other as the squares of the integers, or 1:4:9:16, and so on.

793. Mechanical analogy. A model of the rotating electron of the Bohr atom would be a ball rolling around in a horizontal circular groove part way down the vertical shaft of a mine. Its kinetic energy would of course be positive, but its potential energy with reference to the surface of the earth would be negative. This negative energy would be progressively increased if the level of its orbit were brought closer and closer to the bottom of the pit, while its kinetic energy must be supposed to increase also, assuming a higher rotational velocity at levels nearer the bottom.

In the same way the total energy of the electron, $-(Ze)e/2r$, increases as r decreases, and work is required to raise it to an outer orbit having less negative potential energy, and where its kinetic energy is also proportionately less.

Suppose we represent the kinetic energy at any level by w ; then since the potential energy is minus twice the kinetic, the total is given by $W = -2w + w = -w$. Let w be 6 in a particular orbit; then the potential energy is -12 , and $W_1 = -12 + 6 = -6$. In an orbit more distant from the nucleus we may suppose $w = 5$; then $W_2 = -10 + 5 = -5$. The difference in energy ($W_2 - W_1$) between the two levels is then $-5 + 6 = +1$, which means that one positive unit of energy would be emitted when an electron shifts from an outer to the next inner orbit, or that the same amount is required to lift it back again, very much as would be the case with the ball rolling in a groove around a circular pit.

794. The "orbits." Although the idea of astronomical orbits in the atom is no longer accepted, the energy levels which they represent still remain as necessary concepts. Hence the word orbit, or ring,

should be regarded merely as a convenient term for something which determines the energy of the electron, but which cannot be expressed in language simpler than the mathematical symbols of wave mechanics.

Even if we accept the orbital hypothesis, it is necessary to use elliptical orbits of varying eccentricity whose planes might be inclined to each other at various angles. The word *shell* is therefore more suitable than "orbit" to describe an energy level. But in the very much simplified discussion that follows, we shall imagine all orbits circular and lying in the same plane. Fortunately, such a simplification yields surprisingly accurate results in the case of the spectrum of hydrogen, and some of the broader features of the spectra of heavier elements. But it fails in an attempt to account for what is known as the "fine structure" of spectral lines.

The innermost orbit of any atom is known as the *K* ring, shown in Fig. 35.† In the case of hydrogen, it contains one electron which,

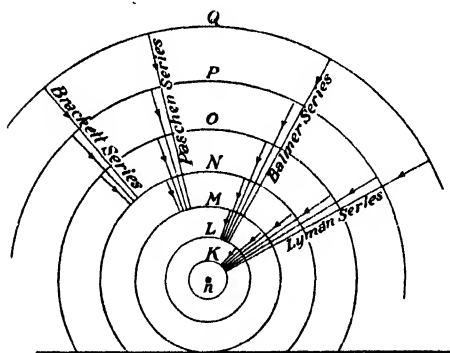


Fig. 35.

however, may be lifted to higher energy levels (orbits) when it absorbs energy. Helium has two electrons, normally in the same ring. But when we come to lithium, the third electron occupies the next or *L* ring, while the *K* ring retains the other two electrons, and never has any more throughout the whole range of known atoms. Beryllium, the next atom, has two

electrons in the *L* ring, and so on through boron, carbon, nitrogen, oxygen, fluorine, and neon, each adding an electron in the *L* ring until, with neon, it has eight. After this the *M* ring makes its appearance, with sodium having one electron normally at that level. This process continues through argon, after which potassium starts the *N* ring with one electron. Its structure is shown diagrammatically in Fig. 36. Then calcium adds one more in the *N* ring. After this the *M* ring builds up to eighteen. During this growth

† This diagram is not drawn to scale. Actually the rings are more widely spaced as we recede from the nucleus, because their radii vary as n^2 , as pointed out in Article 792.

of the *M* ring, the ferromagnetic elements, iron, cobalt, and nickel, make their appearance. When we arrive at copper, we find eighteen electrons in the *M* ring and only one in the *N* ring. Zinc adds another, and so on, until, with the inert gas krypton, it acquires an eighth. The *O* ring next appears and acquires two electrons with ytterbium. Then comes a new disturbance and the *N* orbit builds up to eighteen with palladium. The next element, silver, has eighteen electrons in the *N* ring and only one in the *O* ring. This latter fills up to eight with xenon, and the *P* ring begins with cesium. After cesium there are further irregularities, like those referred to above, which finally, when gold is reached, result in thirty-two electrons in the *N* ring, eighteen in the *O* ring, and one in the *P* ring. The *P* ring reaches eight electrons when we come to radon. Then the *Q* ring begins and reaches two electrons with radium. After this the *P* ring fills up to twelve when uranium is reached. This heaviest element has electrons distributed as follows, beginning with the *K* ring:

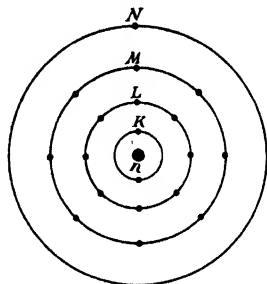


Fig. 36.

$$2 + 8 + 18 + 32 + 18 + 12 + 2 = 92.$$

795. Meaning of the rings. It should be noticed that the two inner and two outer rings never exceed the numbers obtained during their first formation, while the *M* and *O* rings each compound once, and the *N* ring compounds twice. Also, at no time does the outer ring contain more than eight electrons. Each time it reaches that number, the element is one of the inert gases, neon, argon, krypton, xenon, or radon. This seems to indicate a kind of saturation when eight electrons fill the outer orbit, and it is associated with high chemical stability and zero valence. This property of the number eight extends even to compounds when there is a union between elements whose outer-ring electrons add up to eight, or a multiple of eight. Such compounds are far less chemically active than others.

If there is only one outer-ring electron, the atom has a *positive* valence of 1. That is, it may unite, in a diatomic compound like HCl, with another atom of single *negative* valence having seven electrons in the outer ring, or one less than the inert eight. The atom having a single electron in its outer ring must lose that electron to become a positive ion, while the other atom of seven outer-ring

electrons must pick up an electron to become a negative ion. Thus a union of these ions yields eight outer electrons. Similarly two outer-ring electrons indicate a positive valence of 2, while six indicate the same negative valence. An atom like nitrogen, having five outer-ring electrons, may have a valence of 5 or 3, as is well known. Thus we see that the chemical properties of an element depend upon the outer ring only. Many of its physical properties probably do also, such as color, but the entire array of orbital electrons is concerned with the production of the spectrum and other optical properties of an element.

796. The production of the spectrum. In general, when an electron changes from one possible energy level to another lower one, it emits a quantum of energy of definite frequency, and so produces a line of its spectrum associated with the particular level to which it falls. Thus the so-called *K* spectrum is produced by electrons shifting to that ring from rings at higher levels, to which they must have been temporarily lifted when the atom absorbed energy in order to emit. This *K* series for hydrogen is the Lyman series already discussed, and is in the ultra-violet. But with increasingly heavy atoms, it shifts progressively toward still shorter wave lengths, and so passes over into the region of the very short waves belonging to X-rays. This speeding up of the frequency would be expected from the greatly increased energy due to the heavier nucleus, so that a change from the *L* to the *K* level would liberate more energy from platinum than from hydrogen, for instance. Since $W_2 - W_1 = h\nu$, the frequency is greater when a greater amount of energy is liberated.

Beyond nickel (number 28) the *K* spectra all have the same characteristics. Two of the four principal lines are very close together

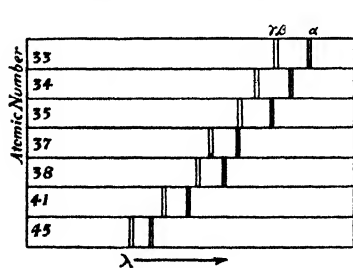


Fig. 37.

(α_1 and α_2), but one is twice as intense as the other. The two others (β and γ) are farther apart and at some distance from the α lines.

If these spectra are arranged one below the other so that equal wave lengths correspond, they may be compared with respect to their atomic numbers, as in Fig. 37.

This remarkable progression was

discovered by Moseley† in 1913. It shows a steady decrease in the wave length for the same line of the series, as the atomic number

† H. G. J. Moseley, an English physicist of unusual ability, killed in the World War at the age of twenty-seven.

increases. Moseley investigated the elements from aluminum (number 13) to gold (number 79), and formulated the remarkable law

$$\nu = R(Z - \alpha)^2,$$

where Z is the atomic number, and R and α are constants. Thus the frequency ν increases with the square of the atomic number. Later observations have extended the range down to beryllium (number 4) and up to uranium (number 92), and the added elements continue to verify Moseley's law very nearly.

A similar relation exists in the L spectrum, which is obtained when electrons fall into the L ring from which they have been lifted. But as we should expect, these spectra have lower frequencies because the greater distance from the nucleus involves smaller energy changes, and it is possible to obtain such X-ray spectra only from chromium (number 24) up through uranium (number 92). The M and N spectra have of course still longer wave lengths, and the former, using X-rays, has been observed only for elements above dysprosium (number 66). In general, the visible spectra are associated with the outer rings. In the case of the lighter elements, they belong to the L and M series. For hydrogen, as shown in Fig. 35, the K or Lyman series is in the ultraviolet, the L or Balmer series is visible, and the M or Paschen series is in the infrared. An N group of lines, known as the Brackett series, of very long wave length, has also been found, as well as an O group, known as the Pfund series, of even longer wave length.

From what has been said it is evident that the atomic number as determined by Moseley's law is more significant in locating an element in the series of elements than the atomic weight, and justifies the reversal of the serial order in the four cases where the two modes of classification were in conflict.

797. Wave mechanics. As has been pointed out, the Bohr atom, although an important stepping stone to further knowledge, has been superseded by a new conception of matter based upon the so-called **wave mechanics**. This theory aims to harmonize the apparent contradiction between the classical wave theory of radiation and the quantum theory. The former represents radiation as a continuous flow of energy in waves, and is needed to explain diffraction, polarization, and so forth, while the latter is corpuscular and is needed in the theory of the production and absorption of radiant energy.

Wave mechanics was developed during the years 1922-27, first by deBroglie,† in Paris, and then by Schrödinger,†† in Zürich. Its

† Louis deBroglie, a French mathematical physicist.

†† Erwin Schrödinger, born in Austria, now professor in the University of Graz.

very complex theory lies beyond the scope of this book, but an idea of its chief concepts may be obtained from the following considerations.

Let us consider a group of electromagnetic waves moving together through space, like the bow waves of a steamer moving over calm water. The energy of such a group, according to the quantum theory, is $h\nu$. Its equivalent mass is $h\nu/c^2$, as was shown in Article 307, and its equivalent momentum is the product of this mass and the velocity of light, or $h\nu/c$. Suppose such a group of waves is associated with an electron, or perhaps we should say *is* the electron. If the mass of the electron is m , and its velocity is v , its momentum is mv . Then, assuming that it owes its momentum to the energy of the group of waves associated with it, we may write

$$mv = h\nu/c. \quad (1)$$

But $c = \lambda\nu$; therefore, substituting for c in (1) and solving for λ , we obtain

$$\lambda = h/mv. \quad (2)$$

This fundamental postulate of wave mechanics gives us the "deBroglie wave length" of a moving electron. It tells us that the wave length of such a moving group of waves varies inversely as the group velocity and inversely as their mass equivalent, or the mass of the electron with which they are associated. Thus high-speed electrons should consist of very short waves.

798. Electron waves. In 1927, Davisson and Germer, working in the Bell Telephone Laboratories, made the important discovery that free electrons are endowed with the wave properties assumed in the preceding article. They showed that a stream of electrons striking a crystal is reflected or diffracted in the same manner as X-rays. If the velocity of the electrons is constant, they are diffracted at a certain angle of incidence that depends upon this velocity, and the maximum intensity of the diffracted beam occurs at an angle of diffraction equal to the angle of incidence. Other velocities call for other angles of incidence, and again the maximum intensity is with those diffracted at an angle equal to the new angle of incidence. This would all be true of X-rays of different wave lengths instead of different velocities.

It thus appears that the velocity of the incident beam of electrons plays the same role as the wave length of X-rays, and Davisson and Germer concluded that the electron is either a bundle of waves or is associated with *waves whose frequency varies with the velocity*. In

this respect the phenomenon differs from X-rays because these have only one velocity, that of light, and the wave frequencies depend upon the radiating source. Moreover, X-rays, unlike electrons, are not deflected by electrostatic or electromagnetic fields.

In order to test the validity of the assumptions contained in equation (2) of the last article, we may calculate the length of electron waves in terms of the potential difference that determines the velocity of the electrons. Since their kinetic energy equals the work done by the electrostatic field that accelerated them, we may write

$$\frac{1}{2}mv^2 = eV, \quad (1)$$

where V , the accelerating potential, and e , the electronic charge, are measured in absolute electrostatic units. But from the preceding equation (2), $v = h/\lambda m$. Substituting this in (1) above gives

$$\frac{mh^2}{2\lambda^2 m^2} = eV.$$

$$\therefore \lambda = \frac{h}{\sqrt{2meV}}, \quad (2)$$

from which λ may be calculated in terms of the potential V .

If V is measured in volts (10^8 e.m.u.), if e is expressed in e.m.u. also, and if λ is measured in angstroms, while h and m are given their accepted values in c.g.s. units, equation (2) reduces approximately to the simple form

$$\lambda = \sqrt{\frac{150}{V}}. \quad (3)$$

Thus 150 electron volts cause a wave length of one angstrom, while 1000 electron volts cause a wave length of the order of 0.4 Å. That means that ordinary-speed electrons have wave lengths comparable with those of fairly hard X-rays.

If a beam of electrons is accelerated by a particular potential difference V_1 , and is then diffracted by the surface of a crystal, we may determine its wave length λ_1 by measuring the angles where the diffracted rays are maximum, on the assumption that the beam behaves like X-rays of the same wave length. Such "observed" values agree remarkably well with the values calculated from equation (3) above, and seem to establish the wave character of electrons.

This hypothesis has been still further strengthened by the experiments of G. P. Thomson, in which beams of electrons were passed through powdered crystals. They produced photographic patterns like those made by X-rays passed through a similar powder. There

is similar evidence that the proton also consists of, or is associated with, waves, when in motion; hence we may be on the eve of finding that there is no ultimate particle of matter in the ordinary sense of the word, and that everything is reducible to wave motion. But waves of what? And how created? These are still unsolved mysteries.

799. Disintegrating the atom. There are certain of the heavier elements, notably radium, which disintegrate spontaneously. This phenomenon, known as radioactivity, will be discussed in the next chapter. Since the discovery of radioactivity in 1896, frequent attempts have been made to bring about both disintegration and synthesis of atoms artificially, and thus realize the alchemist's dream, the transmutation of metals.

There are, broadly speaking, four ways of attacking the problem, all of which consist in bombarding the nucleus of the atom with high-speed corpuscles. We may use as projectiles the "alpha rays" (ionized helium atoms) spontaneously emitted by radium and other active substances. We may use high-speed protons and high-speed deuterons. And finally, we may use neutrons ejected from a beryllium or some other metal target subjected to the impact of high-speed positive corpuscles.

In all of the ways just enumerated, a great number of the elements have been transmuted with the simultaneous emission of corpuscular rays. The lighter elements as far as potassium were the first to yield to this treatment, though helium has so far resisted because of the stability of its nucleus, as stated in Article 787. Carbon and oxygen also show a similar stability, but have been disintegrated by neutron bombardment.

800. Alpha-ray bombardment. The first really successful attempts to disintegrate the nucleus by means of positively charged corpuscles were made by Lord Rutherford of Cambridge University. As the nucleus and the alpha particle used to bombard it both have positive charges, a strong adverse field has to be overcome before a disruptive collision can occur. The kinetic energy needed by the projectile is

of the order of 10^{-5} erg, but this is approximately the measured energy of some of the faster alpha particles. Rutherford's method was to observe the scintillations excited by alpha particles emitted by radium

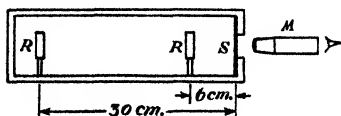


Fig. 38.

C on a screen *S* of zinc blende, as shown in Fig. 38. The radiating source *R* was placed in a tube filled with the stable gas oxygen. The

scintillations on S , observed through the microscope M , occurred whenever an alpha particle struck the screen, but they ceased altogether if R were more than 6 cm from S . Then nitrogen was introduced into the tube, and the scintillations reappeared and persisted with increasing separation between R and S up to 30 cm. These new scintillations were proved to be due to the impact of protons ejected from a nitrogen nucleus by the impact of the alpha particle. As protons have only a quarter the mass of alpha particles, they are faster and more penetrating, and are able to pass through a much greater layer of gas before being absorbed. Rutherford also obtained similar results by bombarding aluminum and other elements of low atomic weight, as stated above.

The process of releasing a proton by alpha-ray bombardment is probably as follows: The alpha particle consisting of two protons and two neutrons enters the nucleus of the bombarded element, and, releasing a proton, remains to transform the nucleus into one having an atomic weight three units higher than before. Also, as two positive charges have been added and only one lost, the result should be an element of the next higher atomic number.

The detection of protons ejected by alpha-ray bombardment may be accomplished in several ways besides the use of a zinc-blende screen. One method is the production of the Wilson water-vapor tracks. Another very valuable device is the **Geiger counter**. This consists essentially of a needle point or fine wire mounted along the axis of a cylindrical ionization chamber. The needle or wire electrode is charged positively to from 1200 to 1500 volts. When an ionizing particle enters the chamber through a small opening, it produces ions in the gas that the chamber contains. These ions cause the electrode to discharge, and an electrometer or other indicator connected to the electrode responds. The number of such responses per second is a measure of the rate at which the ionizing particles are produced.

In a fourth method due to Leprince-Ringuet, a proton or electron enters an ionizing chamber formed by the plates of a charged condenser. The condenser is connected to the grid of a three-element tube so that the grid potential is altered. The resulting effect on the plate current is magnified by several stages of amplification and is then made audible, or it turns a minute mirror which reflects a beam of light. The deviations of this beam may then be photographed on a moving film, so that each captured ion is shown as an abrupt jog in the line traced by the luminous pencil of light.

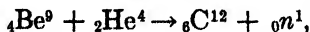
801. Rays from beryllium. Before describing the use of neutrons as projectiles, we shall first give a brief account of their discovery. In 1930 Bothe and Becker, German physicists, experimenting with alpha rays from polonium, found that some of the lighter elements, when exposed to these rays, gave off an extraordinarily penetrating radiation, which they supposed to be electromagnetic, like X-rays. They obtained the best results by bombarding beryllium, though boron and lithium gave out a similar but less intense radiation.

These rays were studied by Mme. Irene Curie-Joliot and her husband, M. F. Joliot, and in 1931 they measured their absorption by lead and found that they had a greater penetrating power than the rays from any known radioactive material.

Those emitted by beryllium actually penetrate through 40 cm of lead. They are thus seen to be very poor ionizers, a fact also known from direct observation of their ionizing power. In testing this latter property, Mme. and M. Joliot found that when these new rays passed through a substance such as paraffin, which contains hydrogen, protons were ejected having a speed of about 3×10^9 cm/sec. This corresponds to a range of 30.4 cm in air, and indicates enormous energy on the part of the ionizing rays when they do succeed in disrupting an atom.

802. Discovery of the neutron. In 1933, Chadwick, of Cambridge University, proved that the penetrating rays from beryllium were material particles having no charge. They owe their remarkable penetrating power to this fact, because they are not deviated by the fields within the atoms, through which they pass without disrupting them, except in the rare instance of a nearly head-on collision with a nucleus. Thus they lose little energy and pass through a great thickness of matter before being stopped. These particles are called **neutrons**. Their atomic weight, at the time of this writing, is taken to be 1.0091, which is slightly greater than that of a proton. As they are uncharged, their atomic number is zero.

The liberation of neutrons from beryllium may be accounted for by the capture of an alpha particle (ionized ${}_2\text{He}^4$) with the formation of the usual carbon atom of atomic weight 12, or ${}_6\text{C}^{12}$, and the emission of a neutron, according to the nuclear reaction



where n means neutron. If the protons ejected from hydrogen by the impact of neutrons have a velocity of 3×10^9 cm/sec., this would demand about the same velocity for the neutron, since their masses

are almost identical. This high corpuscular speed seems reasonable according to Chadwick, although, as they are electrically neutral, the velocity cannot be measured by deviation in magnetic or electrostatic fields, as is that of helions, electrons and protons. At any rate, the recoil velocity of the atoms struck by neutrons indicates a very high speed, and the corpuscular nature of the rays is further indicated by the fact that the emission from beryllium is most penetrating in the direction of the alpha particles that caused it. This would not be the case if the rays were electromagnetic in character.

According to Chadwick, a proton moving with one tenth the velocity of light travels on an average only a foot in air at normal pressure before being stopped by a head-on collision with a nitrogen nucleus. A neutron with the same speed will collide only once in about a thousand feet, and even then may continue for a mile or more before its energy is exhausted. However, when it does strike a nucleus with a head-on collision, the result is highly destructive.

803. The cyclotron. The problem of breaking down the nucleus of an atom depends upon the use of high-speed particles of sufficient mass. Such particles may be protons, deuterons, neutrons, or alpha particles, as we have seen. In order to give a charged particle the necessary speed, high voltages are usually necessary, and much investigation has been directed toward producing them, as in the case of the Van de Graaf generator, described in Article 607. But this difficulty has been eliminated by a very ingenious device invented by Lawrence and Livingston of the University of California.

The **cyclotron**, as it is called, consists of two hollow half-cylinders *A* and *B* (Fig. 39) insulated from each other and enclosing a vacuum chamber. These half-cylinders are subjected to an alternating e.m.f. of from 10,000 to 30,000 volts. The whole arrangement is placed between the poles of an enormous electromagnet capable of producing a field of 16,000 gauss or more, normal to the plane of the diagram, where it is supposed to be directed upward. Then a moving charged particle would follow a circular path in the plane of the diagram, as explained in Article 752.

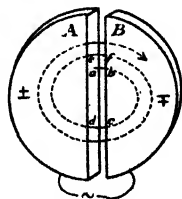
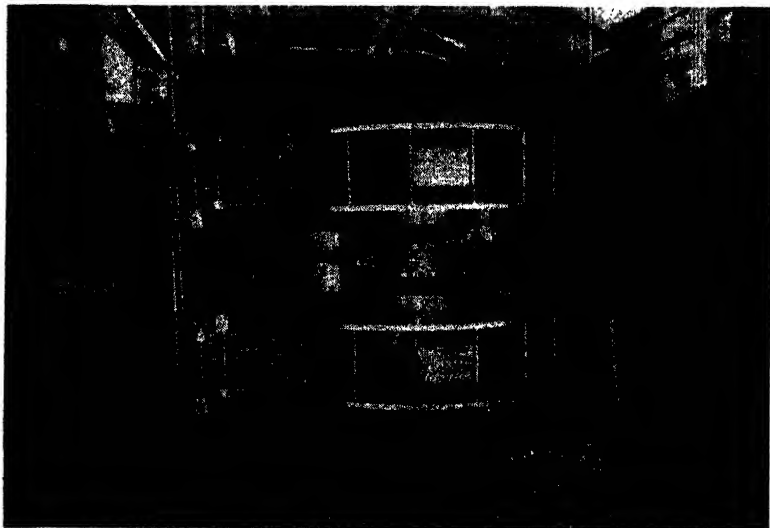


Fig. 39.

Now suppose that deuterons are produced by some means (not indicated in Fig. 39) at the point *a*, and suppose that there is a difference of potential of 10,000 volts between *A* and *B*, with *B* negative and *A* positive. The positively charged deuterons will be drawn across from *a* to *b* into the space within *B*, and acquire considerable

speed in the process. Under the influence of the magnetic field, a deuteron follows a circular path whose radius depends upon the potential and the magnetic field strength. If the magnitude of these quantities and the frequency of the e.m.f. are correctly adjusted, the deuteron reaches c at the moment when the charges on A and B reverse sign. Thus the particle is drawn across from c to d and acquires an increased velocity while making the jump. It now traces



Courtesy Professor DuBridge, Rochester University.

Plate 28.

Photograph of the cyclotron recently installed at Rochester University. Note the massive H-type magnet with vacuum chamber between the beveled pole pieces. This cyclotron is designed to produce high-speed protons rather than deuterons.

a new circular path of larger radius than before, because it is moving faster, as is shown in the case of an electron by equation (2) of Article 752. The velocity acquired in the process of acceleration is given by this equation, and the time required for half a turn is obtained by dividing the half-circumference, πr , by the velocity. Hence

$$t = \frac{\pi r}{v} = \frac{\pi r m}{B r e} = \frac{\pi m}{B e},$$

which shows that the interval between jumps, when new speed is gained, is independent of the radius and therefore constant. Thus with constant frequency and field strength, the accelerating process continues, twice per revolution, and the deuteron travels in a kind of

spiral with rapidly increasing velocity. At the end of 100 "laps," calculation shows that with 10,000 volts applied, an energy equivalent to 2 million electron volts is easily obtained. With 30,000 volts applied and a field of 16,000 gauss, the result may be as high as 5 million electron volts. Finally these high-speed deuterons reach the edge of the circular chamber and are there allowed to escape through a narrow sheet of metal foil. Then they collide with a suitable target, or enter an ionizing chamber containing a gas to be bombarded.

804. Production of neutrons. Neutrons cannot of course be accelerated by an electrostatic field. But high-speed neutrons may be obtained from the impact with a metal target of deuterons produced in the cyclotron. This breaks them down into protons and neutrons. Then a lead screen absorbs the protons, leaving only the neutrons, which pass through almost unimpeded.

The method for producing neutrons described in Article 801 is particularly effective when the beryllium is bombarded by a stream of deuterons instead of alpha particles. The result is an isotope of boron and an ejected neutron, n , according to the nuclear reaction, ${}_4\text{Be}^9 + {}_1\text{H}^2 \rightarrow {}_5\text{B}^{10} + {}_0n^1$. The energy lost in the disintegration of the ${}_1\text{H}^2$ nucleus (or deuteron) gives the neutron great speed. It is therefore not necessary to make use of the cyclotron in order to obtain sufficiently high-speed deuterons. Moderately strong electrostatic fields give the deuterons speed enough to obtain a plentiful supply of fast neutrons from beryllium.

805. Composition of the nucleus. With the exception of hydrogen, the atomic weights w of the elements are equal to, or greater than, twice their atomic numbers Z , or $w \geq 2Z$. This inequality increases as we ascend the scale of the elements, and implies an increasing number of neutrons combined with a proportionally smaller number of protons. In general, the nuclear composition may be expressed by $lp + mn$, where l and m are whole numbers, and m is equal to, or greater than, l . The value of l increases from unity with hydrogen, to 92 with uranium, and m varies from 0 to 146. Thus in one bismuth isotope of atomic number 83, $l = 83$, and $m = 126$; so we may express the nuclear structure by $83p + 126n$. Then taking the atomic weights of both p and n as practically unity, we obtain $83 + 126 = 209$, as the atomic weight of the nucleus.

It is a curious fact that elements of *even* atomic numbers have in general more isotopes than those having *odd* atomic numbers. Another curious fact is associated with the **mass number** of an isotope. This is defined as the integer nearest the atomic weight of the isotope.

It has been observed that, among the lighter elements at any rate, the isotopes whose mass numbers are even, are more common and more stable than the isotopes whose mass numbers are odd. That is, they are less easily disintegrated by corpuscular bombardment. These facts may be related to the following peculiarities of the series of elements. Above nitrogen, ${}^7\text{N}^{14}$, if we take the mass numbers of what may be called the typical isotope of each, we find that with very few exceptions they are alternately odd and even, according to whether the atomic numbers are respectively odd or even. Usually the atomic weights advance one unit in going from odd to even values, and three, five, or occasionally seven units in going from even to odd. These longer steps suggest the formation of a less stable and therefore less abundant atom than those formed by the short step, as is apparently the case.

806. The packing effect. Careful measurements by means of the mass spectrograph, especially with the type used by Bainbridge, reveal the fact that the *isotopic weights* are in general not exactly whole numbers, as was at first supposed, but that they are more nearly so than the atomic weights of the composite elements made up of groups of isotopes. The difference between the isotopic weight W and the mass number M is due to what is known as the **packing effect**, caused by loss of mass in the formation of the element. If this difference is divided by the mass number, we obtain the so-called **packing fraction**, $(W - M)/M$, which measures the proportional variation of mass from the integral value.

As the mass numbers are all based on oxygen arbitrarily taken as 16, with that element, $W = M$, and the packing fraction is zero. But lighter elements have positive packing fractions, as their isotopic weights are all larger than their mass numbers. Starting with a value for hydrogen of $(1.0081 - 1)/1 = 0.0081$, the packing fractions decrease rapidly until, with oxygen and its neighbors, fluorine and neon, it passes through zero to become negative. With increasing mass numbers, M is greater than W , and the packing fraction is negative, reaching a minimum value of about -0.001 with chromium and nickel. Then it rises again, and near osmium (${}^{76}\text{Os}^{191}$) becomes slightly positive and continues gradually increasing through the remaining heavy and largely radioactive elements.

If hydrogen were taken as the basis of the packing effect instead of oxygen, then the packing fraction of hydrogen would be zero, and all the other elements would have negative values. As the hydrogen nucleus (proton) is a basic unit in building up heavier atoms, the

negative packing fractions of the other elements with respect to hydrogen mean that, like helium, they have all lost mass and radiated energy in the process of forming a stable unit.

The mass lost in the process of forming an atom from its constituent parts is called by Aston the **mass defect** and it is really more significant than the packing fraction. In Article 787 we have already calculated the mass defect, as a loss of atomic weight, for the helium nucleus. The calculated value for any atom is based on an assumed nuclear structure containing ultimately only protons and neutrons. If this assumption is correct, the number of protons must equal the atomic number, and enough neutrons must be added to make up the mass number. To this must be added the mass of as many outer electrons as the atomic number. Thus we might regard the chlorine isotope, ${}_{17}\text{Cl}^{35}$, as made up according to the formula $17p + 18n + 17e$. If the actual mass of an atom is deducted from its mass calculated in this way, the result is the mass defect.

SUPPLEMENTARY READING

H. A. Wilson, *The Mysteries of the Atom*, D. VanNostrand, 1934.

W. Aston, *Mass-Spectra and Isotopes*, Edwin Arnold, London, 1933.

F. A. Lindermann, *The Physical Significance of the Quantum Theory*, Clarendon Press, Oxford, 1932.

D. E. Richmond, *The Dilemma of Modern Physics (Waves or Particles?)* G. P. Putnam's Sons, 1925.

G. E. M. Jauncey, *Modern Physics*, D. VanNostrand, 1932.

F. Rasetti, *Elements of Nuclear Physics*, Prentice-Hall, 1936.

PROBLEMS

1. Using equation (3) of Article 798, calculate the electron wave length in a field of 200 volts. *Ans.* 0.866 \AA .

*2. Taking the mass of the hydrogen atom as $1.662 \times 10^{-24} \text{ g}$, and its atomic weight as 1.0081, calculate the mass defect of the isotope of copper, ${}_{29}\text{Cu}^{63}$. *Ans.* $8.38 \times 10^{-25} \text{ g}$.

CHAPTER 59

Radioactivity

807. The Becquerel rays. Immediately after the discovery of X-rays by Röntgen in 1895, Henri Becquerel, a French physicist, began investigating a possible connection between fluorescence and the new form of radiation. He experimented with various fluorescent substances to see if they would fog a photographic plate in the dark, after having been exposed to sunlight. Among these he tried crystals of uranium salts, and found that they did emit penetrating rays like those discovered by Röntgen, but much to his surprise, these were emitted just as well without previous exposure to sunlight.

As soon as this remarkable fact was made known, in 1896, numerous other investigators began looking for a similar result with other substances, and thus the salts of thorium were found to possess the same unexplained property.

About this time it was discovered that these radiations, as well as X-rays, were able to ionize a gas so as to make it conducting. This important effect proved of the greatest value in the study of the new radiations, because it offered an excessively sensitive test for the presence of any radioactive substance in quantities too small for chemical analysis.

808. The discovery of radium. Becquerel soon found that rays from compounds of uranium were emitted regardless of the nature of the compound, and that those which were not fluorescent were active as well as the fluorescent ones. This meant that the property was inherent in the element itself, being independent of its chemical combinations. Following these discoveries, Mme. Curie, working in a Paris laboratory, began an exhaustive examination of these phenomena. She used especially the mineral pitchblende obtained from the mines in Joachimsthal, a town in Czechoslovakia. This ore contained from 70 to 75 per cent of a black uranium oxide, but to her surprise she found the ore three or four times more active than could be accounted for by the amount of uranium present. This led her to suspect a new and much more radioactive substance present in minute quantities, and she undertook its separation with the help of her husband, Pierre Curie.

The process followed was that of fractional crystallization in conjunction with tests for radioactivity, through the ionization of air, as detected with an electroscope. This process was necessary in order to separate the new substance, radium, from barium, which is also contained in the ore, and which closely resembles radium in chemical behavior. Barium chloride, however, has a slightly different solubility from radium chloride, and when a mixture of these two crystallizes out of a solution, the first crystals to form are richer in the radium salt than the later ones. So by selecting these more active crystals, dissolving and recrystallizing them, a stronger and stronger product was gradually obtained. At length, in July 1898, the Curies announced a new radioactive "substance" that Mme. Curie named polonium after her native Poland. Then, on December 26, 1898, came the announcement of an extremely active element that they named radium, but it took nearly four more years to obtain a concentrated radium chloride. The isolation of the pure element radium, not combined in a salt, was later also achieved by Mme. Curie.

809. Properties of radioactivity. The discovery of radium was followed by that of actinium, another highly radioactive element, which Debierne announced in 1899. Then began the period of investigating the physical behavior of these strange new substances. Among the most successful in this field were Rutherford, Soddy, and Fajans. They soon found that the "rays" were even more complex than was at first supposed. In the presence of a transverse magnetic field some were bent one way, some another, while others could not be deviated at all, even in the strongest fields obtainable. The direction of bending showed that the kind most easily affected were negatively charged, the less easily bent were positively charged, and the others carried no charge at all, but all three could ionize a gas, and fog a photographic plate. These rays are known as beta, alpha, and gamma rays, respectively, and their relative paths in a magnetic field acting perpendicularly to the plane of the paper are indicated in Fig. 40.

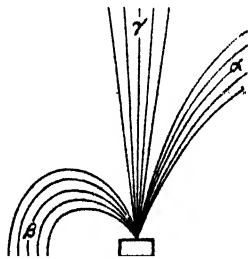


Fig. 40.

Another important physical property of radioactive substances is that they are warmer than their surroundings. This is due to the stoppage of the rays by the material which emits them, causing an evolution of heat. Most of this is due to the alpha particles, whose

energy, as we shall see, is much greater than that of the other rays. In fact, these rays are responsible for nine tenths of the total energy emitted by radium. The amount of heat thus evolved by the alpha rays of a single radioactive substance, assuming all of them to be absorbed within its own mass, is equivalent to the total kinetic energy, or

$$W_K = \frac{1}{2}nmv^2 + \frac{1}{2}nm^2v^2/M,$$

where n is the number of particles of mass m ejected per second, and M is the mass of the atom from which they were thrown. The second term of the equation allows for the energy of recoil of the nucleus, and is calculated as follows: The recoil velocity of the atom is obtained (as for a gun firing a projectile) by equating momenta, giving $MV = mv$. Its kinetic energy is $MV^2/2$; therefore, substituting $V = mv/M$, we obtain for a single recoil, $m^2v^2/2M$, or $nm^2v^2/2M$ for n recoils, as above.

If we allow for the various other radioactive substances present in radium, this calculation gives 1.45×10^6 ergs per second per gram, or 124.8 gram calories per hour, per gram of the substance. Because of the additional heat due to the beta and gamma rays, amounting to 12.6 per cent of the heat just calculated, this value must be increased to 140.1 calories per hour, which agrees fairly well with an observed value of 132 obtained by Duane.

Another important property of radioactive substances is their ability to cause certain bodies to fluoresce or phosphoresce, the latter involving luminosity after the exciting source has been removed. Zinc blende is particularly sensitive to alpha rays, and screens coated with it show minute scintillations wherever it is struck by the particles. This substance is much used in making luminous paints, in connection with very small quantities of radium which is not luminous itself.

The mineral willemite is also fluorescent under the action of all three kinds of rays, while barium platinocyanide is most sensitive to the beta and gamma rays. Kunzite, zinc blende, and fluor spar exhibit phosphorescence under the action of beta rays, the luminosity persisting for some time after the bombardment has ceased.

Radioactivity produces a discoloration of certain substances, such as glass, which turns brown or violet after long exposure. Diamond and kunzite become green, but this color can be removed by heating, and is closely associated with luminescence.

Chemical and physiological effects are also produced by radioactivity. Radium decomposes water, causes oxidation of certain

substances, produces synthesis among others, and affects a photographic plate like light. The physiological effect is associated with ionization, and the beta and gamma rays are the more effective in this respect because of their greater penetration. These effects result in destroying certain living cells, bacteria, and so on, and in producing chemical and structural changes in the tissues.

810. The alpha rays. By means of methods similar to those used in the study of canal rays, the alpha rays were found to be positively charged helium nuclei moving with an initial velocity which, in the case of those produced by radium C', reaches $1/16$ that of light. Alpha rays from other substances have somewhat different speeds, though of the same order of magnitude. The velocity just referred to would be that of the atoms of helium gas raised to a temperature of 7 million degrees centigrade.

The atomic weight of the alpha particle was found to be 4 and in 1903 its identity with the doubly charged helium ion (that is, nucleus) was established by Sir William Ramsay and Professor Soddy, who examined the spectrum obtained from an exhausted chamber in which alpha rays had penetrated, and found it to be that of helium. These rays are powerful ionizers of a gas, and are therefore rapidly absorbed. Those emitted by radium are absorbed in 3.389 cm of air at 15°C and 760 mm pressure. Those from radium C' have a range of 6.971 cm in air, and the average alpha particle from radium C' produces 220,000 ions during its flight.

811. The beta rays. The more easily bent rays produced by radioactive substances are shown to be particles having the negative elementary charge, and as their ratio of charge to mass is nearly the same as that of cathode rays, they are evidently electrons moving with a very high speed approaching that of light. This, as we have seen, results in an increased mass, which makes the ratio e/m smaller than for cathode rays, since e is always the same. These velocities range from 9×10^9 cm/sec. to about 99.8 per cent of the velocity of light, which is much faster than the velocity of the cathode rays in an ordinary X-ray tube, though they may be approached by using extremely high potentials. At 80,000 volts, for instance, the electron speed is only half that of light. With a million volts or more, the speed of cathode rays becomes similar to that of the slower beta rays. But 8 million volts would be needed to duplicate the fastest beta rays. Thus we see that the beta rays have a wide range of velocities, all of which are higher than those of the alpha particles. Therefore, with greater deviability in a magnetic field due to smaller mass, the beta

beam may be spread over a much wider area than the alpha beam, as was indicated in Fig. 40. The most deviable portion of the beam is of course due to the slower moving electrons, whose speed is comparable to ordinary cathode rays.

As a result of their high speed and small mass, the beta particles have much more penetration than the alpha rays. They travel 100 times as far in air before ceasing to ionize, and can pierce even 1 mm of lead, while alpha rays cannot get through a layer of aluminum foil 0.06 mm thick. This means that they are also less readily absorbed by gases, and are correspondingly less powerful ionizers. The total number of ions produced in air by beta rays from 1 g of radium is about 9×10^{14} per second. The gamma rays under the same conditions produce 50 per cent more ions, but the alpha rays yield 2.56×10^{16} ions, or 28 times as many as the beta rays.

812. The gamma rays. These are identical with X-rays, except that they have a shorter wave length and higher penetration. A sheet of lead a few millimeters thick will stop ordinary X-rays, whereas gamma rays can pass through a block of lead 8 inches thick. The shortest gamma wave length yet measured by crystal gratings is produced by radium C, where $\lambda = 0.006 \text{ \AA}$ or $6 \times 10^{-11} \text{ cm}$. This is 10 times shorter than the wave length of the "hardest" X-rays used in making photographs, and was measured by using rock salt as a grating with an angle of $44'$ between the rays and the surface of the crystal. The gamma rays are usually found in conjunction with the emission of a beta particle, and in general both come from the nucleus of the atom, as will be explained later. If gamma rays are used to excite electronic emission from the elements in the same manner as X-rays, their energy and wave length may be obtained from Einstein's equation, $mv^2/2 = eV = h\nu - w$, where $mv^2/2$ is the kinetic energy of the ejected electron, and w is the work necessary to get it out of the atom.

It is quite certain that the high-frequency gamma rays originate in the nucleus, as there are no extranuclear energy levels, even in heavy atoms, which can account for the large energy quantum needed. The simultaneous emission of a beta particle from the nucleus is also significant. Gamma rays have a series of distinct values like the lines in a "characteristic" X-ray spectrum. These values are associated with observed energy differences of the alpha rays, and suggest energy levels *within* the nucleus and the creation of gamma rays by a mechanism similar to that by which X-rays are produced outside of the nucleus.

813. Atomic disintegration. During the first few years after the discovery of radium, the investigators in this field studied the rays and their behavior, as well as the physical and chemical properties of the various radioactive substances. But they were unable to explain the origin and meaning of the rays, or the continuous production of heat by these bodies. Nor could they account for the fact that the heat evolved is independent of their temperature, though it decreases with time.

In 1902 Rutherford and Soddy advanced the theory that the rays and heat developed spontaneously by certain substances were due to disintegration of the atom. The following year they carried this theory still further, suggesting that the atoms representing a definite fraction of the total mass become unstable at a given time, and break up with explosive violence, emitting the various rays and so producing heat. When this explosion results in the expulsion of an alpha particle of atomic weight 4, the atom loses atomic weight by the same amount, thus creating another element, which in turn becomes unstable, and so continues the process.

According to this view, now universally accepted, the atom of a parent substance like radium undergoes a succession of changes as it disintegrates. Some of these result in a decrease in atomic weight, while others that involve only the loss of a beta particle (accompanied by gamma rays) result in no appreciable loss of mass, but cause a change in the chemical properties of the atom. The rate at which these transformations occur differs for different substances, and depends obviously upon the probability that a given atom will become unstable at a given time. This is extremely small in the case of radium, so that it takes 1690 years for a mass of radium to lose half its initial activity by disintegration. Thus the chance that a particular atom will explode during any specified second is vanishingly small. But in a milligram of radium there are 26.6×10^{17} atoms, so that even with so small a probability, many atoms are transformed every second.

814. Rate of decay. In order to determine the rate of decay of a radioactive substance, we assume that the activity of a given quantity is constant, or, what is the same thing, that the probability of disintegration is the same for every atom. Then the activity (that is, rate of decay) of a group of atoms must depend upon their number and must decrease as they disintegrate. If n represents the number of atoms in a given mass, their rate of decay is $-dn/dt$, and this rate, by hypothesis, is proportional to n . That is, $-dn/dt = \lambda n$, where λ is

the constant of proportionality called the **decay constant**. This is a differential equation whose solution is

$$n_t = n_0 e^{-\lambda t}, \quad (1)$$

where n_0 is the initial number of atoms, n_t is the number after a lapse of time t , and e is the base of the Napierian system of logarithms. As the activity a is proportional to n , equation (1) may be expressed by

$$a_t = a_0 e^{-\lambda t}, \quad (2)$$

where the value of λ depends upon the particular substance examined. The graph of this equation is a logarithmic curve, shown in Fig. 41. It reaches zero only when t is infinite, but it falls to half its initial

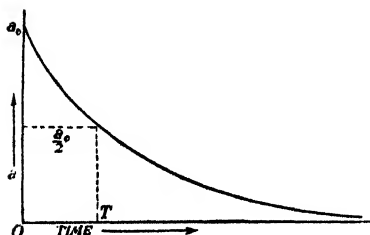


Fig. 41.

value when $t = (\log_e 2)/\lambda$. This particular time, denoted by T , is known as the "half-value period" of the substance, and serves as a useful measure of radioactivity. As was stated above, T for radium has thus been found to be 1690 years, but some other substances decay much more rapidly. The gas radon, which is the first disintegration

product of radium, has a half-value period of only 3.8 days, while radium A, the second disintegration product, decays to half-value in three minutes. The faster a substance decays, the greater its activity, so that weight for weight, radon is enormously more active than its parent radium, while radium A is much more active still.

815. Radioactive energy. The unit of radioactivity is taken as that possessed by the amount of radon in equilibrium with one gram of radium. It is called the curie, and is subdivided into milli-, micro-, and millimicro-curies. This quantity of radon has a volume of 0.63 mm^3 under standard conditions, and weighs $6 \times 10^{-6} \text{ mg}$.

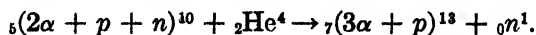
But the activity of such substances may be considered from a different standpoint. The ejected particles and gamma rays proceed from the nucleus, which is the seat of an enormous amount of energy, some of which is released when the rays are emitted. The total energy W obtainable from the complete decay of a radioactive substance may be shown to be given by $W = q_0/\lambda$, where q_0 is the initial rate of heat production corresponding to n_0 . Since a gram of radium develops heat at the rate of 132 calories per hour, or about 0.037 calories per second, we may calculate W from the known value of λ , which is

1.30×10^{-11} reciprocal seconds. The result is 2.8×10^9 gram calories per gram, equivalent to about 8.7 billion foot-pounds! This is by no means the total energy represented by a gram of matter, but it is the heat energy released by a gram of radium if wholly disintegrated. Disintegration is very different from annihilation, and the radioactive atom does not vanish, but ultimately only breaks down into one of lower atomic weight.

816. Induced radioactivity. Almost from the time radium was discovered, frequent attempts have been made to induce radioactivity in nonactive atoms, but until very recently these efforts have been fruitless. In fact, even the unstable radioactive elements have resisted all attempts either to hasten or retard their decay.

Improved methods of attacking the atom with very high speed corpuscles have opened a way to at least partially achieving the desired induced activity. In February, 1934, Professor F. Joliot and his wife, Irene Joliot-Curie, announced the production of artificial radioactivity in boron, aluminum, and magnesium. Their method was to bombard these elements with alpha particles. This resulted in the emission of positrons, which continued for some time after the bombardment. The induced activity of boron decays to 30 per cent of its initial value in 15 minutes, while aluminum decays still more rapidly.

The Joliot's explained their induced activity by supposing that the element captures the alpha particle. In writing the nuclear reactions they regard alpha particles as individual entities within the nucleus, a not uncommon assumption which was used in our discussion of the nuclear structure of lithium in Article 787. Thus the composition of ${}_{10}\text{B}^{10}$ is written ${}_5(2\alpha + p + n)^{10}$ instead of ${}_5(5p + 5n)^{10}$, in accordance with the simpler method of Article 805. Using the Joliot's notation, the nuclear reaction, when ${}_{10}\text{B}^{10}$ is bombarded with an alpha particle (ionized ${}_2\text{He}^4$), is written



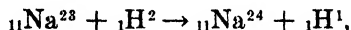
The nitrogen isotope ${}_7(3\alpha + p)^{13}$ is unstable and breaks down into an isotope of carbon, ${}_6\text{C}^{13}$, or ${}_6(3\alpha + n)^{13}$, with the emission of a positron. This latter change would seem to mean a transformation of a proton into a neutron with the loss of a positive electron e^+ , which may be expressed by $p \rightarrow n + e^+$. This tends to bear out the theory that a proton is a union of a positron and a neutron.

Since the experiments described above, Lauritsen, Crane, and Harper, of the California Institute of Technology, have carried out a

suggestion of the Joliot, and bombarded carbon, magnesium, and other substances with deuterons. This resulted in induced radioactivity that was especially pronounced with carbon, whose half-value period was only 10 minutes. Boron was less active and fell to half-activity in 20 minutes.

Examined in a Wilson cloud chamber, the active carbon produced electron tracks, mostly positive, and there was also evidence of gamma-ray emission. The curvature of the positron tracks indicated a wide range of velocities, but the fastest indicated a kinetic energy in agreement with the theory advanced by the Joliot.

Using the great cyclotron of the University of California, Lawrence and Livingston bombarded rock salt with deuterons whose energy was equivalent to more than 2 million electron volts. This resulted in forming an unstable sodium isotope that emits electrons and very penetrating gamma rays. It has a half-value period of 15.5 hours. The nuclear reactions are probably



and



where ${}_{11}\text{N}^{23}$ is ordinary sodium. It captures the deuteron and becomes ${}_{11}\text{N}^{24}$, with the emission of a proton. This unstable sodium then breaks down into magnesium, with the emission of electrons and gamma rays. Radioactive sodium may prove to be of great therapeutic value.

In the University of Rochester, Professor DuBridge is using a recently installed cyclotron to bombard various substances with protons. The results are quite different from those obtained with deuterons, and new cases of induced radioactivity are being discovered there.

Still another method for inducing radioactivity has been very successfully used by Fermi, an Italian physicist, and his school. Their method is to bombard an inactive element with neutrons. A large number of elements have been made active in this way, some of which have quite long half-value periods extending to many hours or even days. Ordinary phosphorus, ${}_{15}\text{P}^{31}$, for instance, when bombarded with neutrons, captures the neutron and becomes the unstable isotope ${}_{15}\text{P}^{32}$, when it emits electrons with a half-value period of 14.5 days.

817. Measurement of radioactivity. The ionizing power of radioactive materials is the basis of most quantitative measurements of

their activity, although the scintillations caused by alpha rays striking a screen of zinc blende are useful in some special cases. Ionizing power is usually measured by means of some form of gold-leaf electroscope, which is one of the most sensitive detectors of feeble ionization currents yet devised. In this way currents of the order of 10^{-15} ampere may be detected. A simple arrangement for comparing ionizing powers is shown in Fig. 42. The electroscope is enclosed in a cylindrical box with glass plates at the ends, on one of which is a scale S by which the deflection of the gold leaf L may be measured. The metal rod supporting the leaf ends on top in a knob, and at the bottom in a plate B within the ionizing chamber. It is carefully insulated by amber bushings AA . The plate C , which carries the radioactive material, is supported by a metal post through which it is grounded. This apparatus is used as follows: A charge sufficient to make the gold leaf diverge is given to the knob. Then the very slow rate at which the leaf moves over the scale as it falls is measured in divisions per minute. The active substance is then introduced and the rate again measured. The difference of these rates is due to the ions created between B and C , which carry off the charge from B to the earth. This rate may then be compared with that caused by a known amount of a standard substance, such as uranium, and the activity obtained in milli-curies.

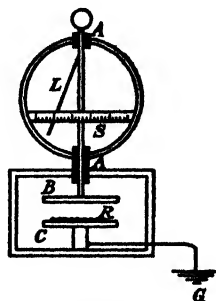


Fig. 42.

818. Radioactive transformation. As has been stated, the loss of an alpha particle means a decrease of four units in the atomic weight of the atom which ejected it. Since the helium nucleus (alpha particle) has two positive charges, this means that the atomic number is diminished by two units, and there must be two less outer electrons remaining after such an explosion. This changes the chemical nature of the atom and gives rise to a new element.

The production of beta rays from the nucleus is a good deal of a puzzle. One theory is that the ejected electron may come from the simultaneous creation of an electron and positron. The positron might then combine with a neutron to form a proton, thus reversing the process, $p \rightarrow n + e^+$, suggested in Article 816. In any case, the nucleus gains one positive charge, but retains practically the same mass as before. This results in increasing the atomic number by one, without change of atomic weight. Two such elements of the

same atomic weight, but of different atomic number and chemical properties, are called **isobars**.

If the expulsion of an alpha particle is followed by that of two beta particles in succession, the resulting atom is an isotope of the former, having an atomic weight lower by four, but with the same atomic number. It is interesting to note that the beta particles seem to go in pairs, giving rise to even atomic numbers. If one is ejected, another must follow, though the emission of a second alpha particle may very briefly retard the emission of the second beta particle. This agrees very well with the nuclear composition proposed in Article 816, since the neutrons also appear in pairs.

819. Radioactive series. There are three distinct series of radioactive substances produced by the disintegration of a parent element. These are the uranium-radium series, the thorium series, and the actinium series. The first is the most important, and we shall consider it in more detail.

Uranium, the parent element, is very much less radioactive than radium, but the two are found together in minerals in the proportions indicated by theory based on assuming one to be a descendant of the other. Therefore there is no doubt that they belong to the same series. The substance which connects the uranium series with that of radium was long unknown, but was finally discovered by Boltwood, of Yale University, who named it ionium. The changes in atomic weight and number that take place in the series as a result of the loss of alpha or beta particles are shown in the diagram of Fig. 43, due to Soddy. The formation of each new substance is effected by the emission of an alpha or beta particle. With radium C there are two possibilities: either a second electron emission forming radium C' followed by an alpha particle to form radium D, or C may form C'', first by the emission of an alpha particle, and then arrive at D by the loss of an electron, thus reversing the order of the events. But the first order, C, C', D is much the more likely, and belongs to the standard type when two electrons are emitted in sequence, whereas the less usual C, C'', D, involves the exceptional case of β , α , β . But the alpha activity of radium C'' has a half-value period of only 1.32 minutes, so that the second beta particle is not long delayed.

Uranium X₁ and ionium are isotopes, as is seen from the diagram; radium B and radium D, radium C and radium E, and radium A, C', and F are also isotopes. Radium F is polonium, discovered by Mme. Curie, and G is lead of atomic weight 206, being one of the isotopes of that element, and found in conjunction with the minerals

from which radium is obtained. Thus the series ends in a stable and fairly common element, much of which must have been formed in this way by the slow process of the disintegration of uranium.

In this series, only uranium, radium, radon, and polonium represent distinct elements. Uranium X_1 is an isotope of thorium (No.

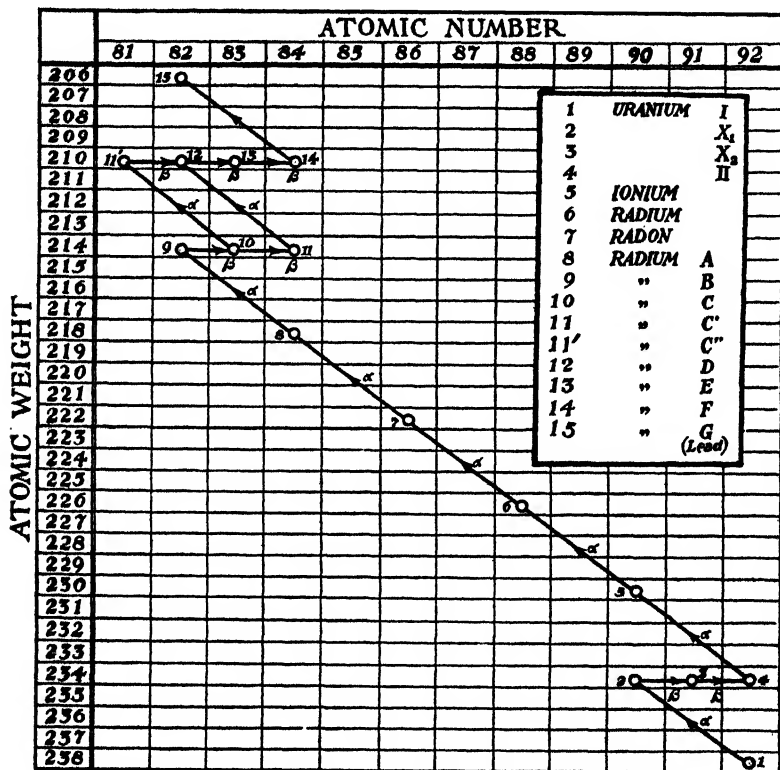


Fig. 43.

90), uranium X_2 is an isotope of protactinium (No. 91), uranium II is an isotope of uranium I (No. 92), radium C and E are isotopes of bismuth (No. 83), and radium C'' is an isotope of thallium (No. 81).

The following table, after Kovarik,[†] gives values of the decay constant λ , half-value periods T , with the nature of the emitted rays of part of the uranium-radium series. The two alternate routes from radium C to D via C' and C'' , respectively, are indicated by the

[†] A. F. Kovarik and L. W. McKeehan, *Radioactivity*, Bulletin No. 51 of the National Research Council, Washington, 1925.

arrows. As to the lives of the slow-decay substances, they are known only approximately, and 1690 years as the half-value period for radium is perhaps too large, although it was originally estimated at 2000 years.

Element	Atomic Number	λ	T	Ray
Uranium I.....	92	4.8	4.6×10^9 years	α
Ionium.....	90	2.96×10^{-13}	7.43×10^4 years	α
Radium.....	88	1.30×10^{-11}	1.69×10^3 years	α
Radon.....	86	2.106×10^{-6}	3.810 days	α
Radium A.....	84	3.85×10^{-2}	3.0 min.	α
Radium B.....	82	4.31×10^{-4}	26.8 min.	β
Radium C.....	83	5.92×10^{-4}	19.5 min.	$\left[\begin{array}{l} \alpha\beta \\ \alpha \\ \beta \end{array} \right]$
Radium C'.....	84	$10^{4.9}$	10^{-6} sec.	
Radium C''.....	81	8.75×10^{-3}	1.32 min.	
Radium D.....	82	1.37×10^{-9}	16.5 years	β
Radium E.....	83	1.65×10^{-6}	4.85 days	β
Radium F (Polonium).....	84	5.886×10^{-8}	136.3 days	α
Radium G (Uranio-Lead)...	82

The two other radioactive series are those of actinium and thorium. The ancestry of actinium is somewhat obscure. According to Rutherford, it may be derived from an isotope of uranium which he calls "actino-uranium," having an atomic weight of 235. This is supposed to change into an isotope of uranium X_1 known as uranium Y, by the emission of an alpha particle. Uranium Y has beta activity and turns into protactinium of atomic number 91 and atomic weight 231. Actinium is formed from the disintegration of protactinium just as radium is formed from ionium, so that actinium and radium belong to the same generation, so to speak, but have different atomic numbers, 89 and 88, respectively. Actinium itself is much more active than radium, having a half-value period of only 20 years. Its descendants are all short lived, the longest half-period, 18.9 days, belonging to radioactinium. The final product is actinium D, which is an isotope of lead, like uranio-lead, but its atomic weight is unknown.

The thorium series is an independent one, starting with thorium and ending with thorio-lead after ten successive transformations. It has one pair of alternative changes like the others already mentioned. Some of its half-value periods are very slow, and some extremely rapid, as is the case with radium. The final product is a lead of

atomic weight 208, which is a different isotope from uranio-lead, whose atomic weight is 206. Thorio-lead is found associated with thorium, and its identity with the final product of the series (thorium D) is accepted.

SUPPLEMENTARY READING

K. Fajans, *Radioactivity*, Dutton, 1922.

J. A. Crowther, *Ions, Electrons and Ionizing Radiations* (Chap. 16), Edwin Arnold, London, 1934.

Rutherford, Chadwick, and Ellis, *Radiations from Radioactive Substances*, Macmillan, 1930.

APPENDIX

The Solution of Problems

The real difficulty in solving physical problems lies in the application of general principles to actual cases. Very often, of course, only a single formula is needed, and if it is already transposed to give the desired quantity in terms of those stated in the problem, there is nothing left but to substitute values and do a little arithmetic. These are the easiest problems, and they have very little intellectual value for the student, though their solution is still worth while because it emphasizes the practical aspects of the theory and develops facility in numerical computation.

But many problems are not so easy. Several equations or principles enter into them, and these must be intelligently combined, or if only one formula is necessary, it may have to be rearranged in order to obtain the result required. In all such cases the student cannot be urged too emphatically to work through to the final solution with algebraic symbols, before introducing their numerical values for the final computation. This not only minimizes the chance of arithmetical blunders, but it is often a great labor-saving device, because what looks like a complicated array of symbols may reduce, as a result of factoring, and so forth, to a very simple expression demanding only a few easy operations.

The next step, after obtaining the simplest possible expression for the unknown quantity, is substituting the given values where they belong. Here there is but one real difficulty, which consists in having the values expressed in suitable and harmonious units. If two velocities, for instance, are given in miles per hour and feet per second, they cannot be added, subtracted, or otherwise combined until reduced to the same units. Similarly, two distances expressed in feet and meters cannot be combined in a single expression until they are reduced to a common measure.

Finally comes the numerical computation. Here, of course, accuracy is absolutely essential, but this does not necessarily mean carrying out a division to five or six decimal places. But one more significant figure at most should appear in the answer to a problem beyond the number supplied by the data. Thus if a distance, roughly measured to the nearest meter, is 97 meters, and a third of this dis-

tance is desired, the answer is 32.3 meters, and not 32.3333+. This long decimal would be justified only if we knew that the distance has been measured correctly to a fraction of a millimeter. Otherwise the added figures are not only unnecessary but actually wrong, because they are misleading.

Many persons believe that while it may be allowable to drop figures to the right of the decimal point in extracting a square root, or as a result of division, and so forth, it is never justifiable to drop figures on the left. This is also erroneous. If a third of the 97-meter measurement is to be expressed in millimeters, it is not 32,333 mm, but 32,300 mm. Similarly, if 97 meters were reduced to inches by multiplying by 39.37, the answer is not 3818.89 inches unless we are sure that the distance is *exactly* 97 meters. If the measurement were only as accurate as assumed above, there may be an error as great as 20 inches, and the distance should be expressed as 3820. In physics, then, it is incorrect to carry out operations further than is justified by the accuracy of the observed data.

As this is not a course in arithmetic, long computations are avoided wherever possible in the problems. The answers given are obtained mainly by using only three or occasionally four significant figures for the ordinary constants and numerical ratios. Thus, except in a few problems where greater precision is obviously necessary, the student may ordinarily use $\pi = 3.14$, $\sqrt{2} = 1.41$, $\sqrt{3} = 1.73$, $g = 980$ cm/sec², or 32.2 ft./sec², one inch = 2.54 cm, one pound = 454 g, and so forth, in spite of the fact that these numbers are known to many more significant figures.

INDEX

(Numbers refer to pages.)

A

Abampere, 595
 Aberration:
 chromatic, 445
 spherical, 389, 411
 Abohm, 602
 Absolute thermometric scale, 175
 Absolute zero, 174
 Absorption of sound, 325, 326, 348
 Absorption, thermal, 261
 Absorptivity, 261
 Abvolt, 603
 Accelerated motion, equations of, 47-49
 Acceleration, 9
 angular, 79
 of falling bodies, 45, 46
 radial, 80
 Accommodation of the eye, 420
 Achromatism, 446
 Acoustics, architectural, 324
 Actinium, 791
 series, 802
 Action and reaction, 37-39
 Adhesion, 117
 Adiabatic curves, 227
 elasticity, 228
 processes, 226
 Afterimages, 512
 Air pump, 134
 Alnico, 670
 Alpha:
 particle, 770
 rays, 793
 Alternating-current generators, 700
 Alternating currents, 702
 Ammeter, 649
 hot-wire, 650
 Ammonia refrigerator, 219
 Ampere, 595
 international, 622
 turns, 671
 Amplitude of harmonic motion, 87
 Anderson's photograph of positron track, 763
 Angle:
 critical, 394
 of contact (capillarity), 147
 of incidence, 287
 of reflection, 287
 of refraction, 294
 of repose, 68
 Ångström unit, 435
 Angular acceleration, 79
 velocity, 79
 Anions, 619
 Anode, 619
 Antinodes, 292
 Arc:
 discharge, 747
 light, 614
 spectrum, 519

Archimedes' principle, 124-125, 136
 Architectural acoustics, 324
 Armature:
 of D.C. generator, 696, 697
 of induction motor, 711
 reactions, 698
 Aspirators, 141
 Astigmatism, 423
 Aston's mass spectrograph, 771
 Astronomical:
 interferometer, 479
 telescope, 427, 428
 Atmospheric pressure, 122
 Atom, 767, 768
 Bohr's, 773
 Atomic:
 disintegration, 782, 795
 models, 772
 number, 768
 "orbits" or "rings," 775-777
 weight, 181
 Atwood's machine, 53, 54
 Audibility, limits of, 323, 324
 Aurora Borealis, 556
 Avogadro's number, 181, 182
 Avogadro's principle, 181, 182

B

Bach's tempered scale, 334
 Bainbridge's mass spectrograph, 772
 Balance, the, 75
 Ballistic galvanometer, 682
 Balmer's series, 520
 Banking of roads, 83
 Bar, unit of pressure, 118
 Barkhausen noises, 674
 Barnett's experiment, 673
 Barometer, 122
 Barye, unit of pressure, 118
 Batteries, 590, 626-635
 Beats between tones, 316
 Becquerel rays, 790
 Bel, unit of loudness, 322
 Bell, electric, 666
 Bells, vibrations of, 344
 Bernoulli's theorem, 140
 Beryllium, rays from, 784
 Beta rays, 793
 Bias, grid, 746
 Binocular:
 prism, 431
 vision, 375
 Biot and Savart's law, 593
 Blake transmitter, 695
 Block and tackle, 73, 74
 Bohr atom, 773, 774
 Bohr's calculation of Rydberg's constant, 520
 Boilers, convection in, 250
 Boiling point, 200, 201, 245
 Bolometer, 612
 Boltzmann's constant, 183

- Boyle's law, 130
 Bradley's determination of velocity of light, 372
 Brewster's law, 489
 Bridge, post office, 656
 slide-wire, 656
 Wheatstone's, 655
 Bridgeman's high-pressure experiments, 197
 Brightness, 365
 British thermal unit, 185
 Brownian movements, 177
 Bunsen photometer, 368
 Buoyancy:
 center of, 127
 of gases, 136
- C
- Cadmium cell, 633
 Caisson, 133
 Calorie, 185
 Calorimeter, Joly's steam, 203
 Camera, 417
 lens rating, 418
 Canal rays, 739
 Candle power, 363
 Capacitance, e.m.u., 657
 e.s.u., 578
 of condensers, 658
 Capacity, specific inductive, 580
 Capillarity, 148
 Carnot's cycle, 231, 233
 Cathode, 619
 ray oscillograph, 736
 rays, 730-736, 738
 Cations, 619
 Cavendish's experiment, 44
 Cell:
 Daniell's, 630
 dry, 633
 Leclanché, 632
 polarization of, 632
 standard cadmium, 633
 storage, 634
 voltaic, 629
 Celsius scale, 160
 Center:
 of buoyancy, 127
 of gravity, 26, 27
 of oscillation, 104
 of percussion, 103
 of pressure, 120
 Centigrade scale, 160
 Central:
 forces, 81, 83
 heating, 251
 Centrifugal:
 reaction, 82
 separator, 85
 Centripetal forces, 82
 C.g.s. system, 5
 Chadwick's discovery of neutron, 784
 Characteristic:
 temperature, 193
 X-rays, 757
 Charge:
 distribution of, 577
 electronic, 731
 energy of, 582
 e.s.u., 563
 surface, 564
- Charles' law, 169
 Chladni's figures, 342
 Chords in music, 330
 Chromatic:
 aberration, 445
 scale, 335
 spectrum, 435
 Chromosphere, 440
 Circuits, electric, 607, 609
 magnetic, 671
 Circular measure, 6
 Clausius' statement of "second law," 234
 Coercive force, 669
 Cohesion, 117
 Coil:
 induction, 687
 spark, 686
 Cold light, 529
 Collision, 112
 Colloid, 241
 Colors:
 addition of, 509
 classification of, 514
 combinations of, 507
 complementary, 508
 of objects, 506
 perception of, 509
 sensitivity to, 511
 surface, 507
 Combining weight, 620
 Comma, interval of, 333
 Complementary colors, 508, 512
 Compound pendulum, 103
 Compressibility, 117
 Compression:
 heat of, 180
 work of, 131
 Compton effect, 758
 Condensation, 199, 202
 by expansion, 210
 Condensers, 578, 657
 capacitance of, 658
 combinations of, 658
 Conductance, electrical, 604
 Conduction of heat, 253
 Conductivity:
 coefficient of thermal, 253
 electrical, 604
 measurement of thermal, 254, 256
 molecular (electric), 624
 tables of thermal, 255-257
 Conductors:
 of electricity, 560
 in electrostatic field, 572
 Conical intensity of light, 364
 Conjugate:
 foci, 402
 planes, 413
 Conservation of energy, 62
 Consonance, 332
 Contact e.m.f., 628, 637
 Convection:
 of heat, 250
 prevention of, 252
 Coolidge tube, 755
 Cooling, Newton's law of, 266
 Coronas, 536
 Corpuscles, 725
 Corpuscular:
 radiations, 757
 theory of light, 370

Corti, arches of, 322
 Cosmic rays, 522, 765
 Coulomb, the, 600
 Coulomb's law (magnetism), 542
 (electrostatics), 562
 Couple, 22
 Critical:
 angle of refraction, 394
 point, 216
 temperature, 216, 217
 Crookes dark space, 727
 Crookes radiometer, 259
 Cryohydrates, 243
 Crystal:
 gratings, 760
 structure, 758
 Crystal detector, 719
 Curie, the, 796
 Curie (J. and P.), discovery of piezo-
 electric effect, 346
 Curie (Mme.), discovery of radium, 790
 Curie-Joliot, discovery of induced radio-
 activity, 797
 Curie point, 552, 553
 Curvature, 7
 effect in surface films, 149
 "Curved ball," 142
 Cycle, Carnot's, 231, 235
 Cyclotron, 785

D

Dalton's law, 210
 Damped vibrations:
 electrical, 715
 mechanical, 296
 Daniell's cell, 630
 D'Arsonval galvanometer, 648
 Davison and Germer's discovery, 780
 Davy, Sir Humphry, test of caloric
 hypothesis, 222
 Debiere, discovery of actinium, 791
 DeBroglie's equation, 780
 Decay of radioactivity, 795
 constant of, 796
 Decibel, 322
 Declination, magnetic, 554
 Defects of the eye, 421
 Degradation of energy, 63
 Density, 5
 current, 604
 measured, 126
 of water, 168
 surface (electrostatic), 572
 table of, 128
 Detectors of radio waves, 719-721
 Deuteron, 769
 Dew:
 formation of, 265
 point, 212
 Dextrorotatory, 501
 Dialysis, 248
 Diamagnetic susceptibility, 664
 Diamagnetism, 552
 Diatonic scale, 329
 Dielectric, 572
 constants, table of, 581
 strength, 580
 Diesis, interval of, 330
 Difference tone, 317

Diffraction, 463-484
 by grating, 481
 by narrow slit, 468
 by perforated screen, 465
 by rectangular aperture, 470
 by straight edge, 469
 by two apertures, 477
 by wire, 468
 Fraunhofer, 473
 Diffusion:
 of gases, 245, 246
 of liquids, 247
 Dimensional formulae, 6
 Diopter, 406
 Dipoles, 548
 Direct-vision spectroscope, 446
 Discharge, arc, 747
 Discharges:
 electrical, 586, 725
 in exhausted tubes, 736-738
 residual, 582
 Disintegrating the atom, 782
 Disintegration of radioactive atoms, 795
 Dispersion, 434
 anomalous, 447
 coefficient of, 444
 irrationality of, 442
 Dispersive power, 444
 Displacement of ships, 126
 Dissonance, 332
 Distillation, 242
 Diving bell, 133
 Dominant triad, 330
 Dominguez's color classification, 514
 Doppler effect, 317, 318, 448, 449
 Double refraction, 489
 vector diagrams of, 493
 wave surfaces of, 490
 Dry cell, 633
 Duane-Hunt relation, 756
 Duane's measurement of heat of radio-
 activity, 792
 Dulong and Petit, law of, 189
 Dyne, 35

E

Ear, human, 320
 Echo, 313
 Eddy currents, 690
 Edison effect, 741
 Efficiency:
 of heat engines, 235
 of luminous sources, 528
 Effusion of gases, 143
 Einstein's photoelectric equation, 749
 Elastic:
 constants, 111, 114
 limit, 109, 112
 Elasticity, 109
 adiabatic, 228
 isothermal, 228
 modulus of, 110
 Electric:
 battery, 590
 circuits, 607-609
 furnaces, 613
 heating, 613
 welding, 614
 Electric current density, 604
 effects of, 591

- Electric current:
 field of, 592, 594, 597
 magnetic action on, 599
 unit of, 595
 Electricity, 558, 559
 e.m.u., 600
 Electrification by friction, 558
 Electrochemical equivalent, 621
 Electrodynamics, 589-605
 Electrodynamometer, 652
 Electrolysis, 617
 Faraday's laws of, 621
 Electrolytes:
 conductivity of, 611, 624
 dissociation of, 617
 Electrolytic:
 reactions, 618, 619
 solution pressure, 629
 Electromagnetic:
 energy, 605
 waves, 717
 Electromagnetism, 661
 Electromagnets, 664-666
 Electrometer, quadrant, 562
 Electromotive force, 602
 of cells, 626
 Electron, 559
 Electron-volt, 735
 Electron waves, 780
 Electrophorus, 583
 Electroplating, 620
 Electroscopes, 561
 Electrostatic (s), 558-587
 charge, energy of, 582
 contour maps, 574
 Coulomb's law of, 562
 e.s.u., 563
 field, 566
 lines of force, 566
 potential, 587
 surface density, 572
 Elevator, problem of, 52
 Emissive power, 262
 Emissivity, 263
 Energy, 58
 conservation of, 62
 degradation of, 63
 intrinsic, 225
 kinetic, 59
 potential, 59
 transformations, 225
 Equilibrant, 18
 Equilibrium:
 conditions of, 23
 of coplanar forces, 24-30
 types of, 64
 Equipotential surfaces, 573
 Equivalent conductivity, 624
 Erg, 57
 Eutectic point, 243
 Evaporation, 199, 200
 cooling by, 203
 Exchanges, Prévost's theory of, 260
 Expansion (thermal), 163
 differential, 164
 coefficient of, 163, 169
 free, 229
 of gases, 168
 of mercury, 165
 surface and volume, 165
 table of coefficients of, 167
 Expansion (thermal) (*Cont.*):
 of water, 167
 work of, 131
 Eye, human, 419
 defects of, 421
 Eyeglasses, 421-424
- F
- Fahrenheit scale, 160
 Falling bodies, motion of, 49, 50
 Farad, the, 658
 Faraday, the, 621, 622
 Faraday:
 dark space, 727
 ice-pail experiment, 565
 laws of electrolysis, 621
 magneto-optical effect, 503
 Fechner's law, 322
 Fermi's neutron bombardment, 798
 Ferromagnetic susceptibility, 654
 Ferromagnetism, 551
 Field, electrostatic, 536
 of induction motor, 709
 Field maps, 571
 Figure of merit, 649
 Films (surface), 146
 between liquids, 152
 colors of, 457
 double, 150
 free, 151
 tension in, 151
 First law of thermodynamics, 225
 Fizeau's determination of velocity of light, 372
 Floating bodies, attraction between, 153
 Flotation, 126, 127
 Flow of liquids, 138
 Fluids, 116
 pressure in, 118
 pressure on, 117
 Fluorescence, 524
 Fluorescent X-rays, 757
 Flux:
 luminous, 364
 magnetic, 662, 663, 678
 Focal length:
 of lenses, 429
 of mirrors, 384
 Foci, sound, 313
 Focus, principal, 384
 Foley's sound photographs, 325
 Foot candle, 364
 Foot-pound, 57
 Foot-poundal, 57
 Force, 17
 and motion, 33
 dimensions of, 36
 moment of, 21
 unit of, 35
 Forces:
 balanced, 18-21
 between currents, 712
 coplanar, 24-26, 28
 Foucault's determination of velocity of light, 373
 Foucault's pendulum, 106
 Fraunhofer:
 diffraction, 473-484
 lines, 440
 Free expansion of gases, 229
 Free vibrations, 295

Freezing:
 by boiling, 216
 mixtures, 244
 point, 194
 point of solutions, 242
 Frequency of harmonic motion, 87
 Fresnel diffraction, 473
 Fresnel's biprism, 455
 Friction, 66
 coefficient of, 67
 problems involving, 68, 69
 rolling, 68
 Fungi, phosphorescent, 526
 Furnaces:
 for central heating, 252
 electric, 613
 Fusion, 194
 change of volume during, 195
 heat of, 197

G

Galileo's observations on falling bodies, 45
 Galileo's telescope, 430
 Galvanic couple, 628
 Galvani's discovery, 628
 Galvanometer:
 ballistic, 682
 D'Arsonval, 648
 fixed coil, 647
 moving coil, 648
 Gamma rays, 522, 794
 Gas law, 175
 Gas thermometer, 173
 Gases, 116
 buoyancy of, 136
 coefficient of expansion of, 169
 compression and expansion of, 131
 diffusion of, 245
 dissolved in liquids, 240
 effusion of, 158
 expansion of, 168
 elasticity of, 228
 free expansion of, 229
 heat of compression of, 180
 ideal, 182
 liquefaction of, 220
 mechanics of, 130-136
 pressure of, 178
 specific heats of, 189, 190, 192
 vapors and, 206-213
 Gauss, the, 663
 Gegenfarben, 513
 Geiger counter, 783
 Generator:
 A.C., 700-702
 compound, 699
 D.C., 695-700
 series, 699
 shunt, 698
 voltage of, 696
 Gilbert, the, 671
 Gilbert's observations on magnetism, 541
 Glaciers, 197
 Governor, flyball, 84
 Gram:
 atom, 182
 equivalent, 621
 molecule, 182
 Grating:
 crystal, 760
 diffraction, 481

Grating (*Cont.*):
 spectra, 482
 Gravitation, Newton's law of, 45
 Gravitational:
 constant, 44, 45
 force units, 46, 47
 waves, 281, 282
 Gravity:
 center of, 26, 27
 specific, 125
 Grid (radio-tube), 743
 Grid bias, 744
 Gridiron pendulum, 164
 "Grounds," 561
 Gyration, radius of, 99
 Gyroscope, 105

H

Hadley's sextant, 382
 Half-period elements, 464, 467
 Half-value period, 796
 Half-wave quartz, 502
 Halos, 536
 Hardness, 114
 Harmonic motion:
 of rotation, 102
 of translation, 101
 simple, 86
 Harmonics, 310-312
 Hasenöhrl's theory of radiation and mass, 269
 Hastings' theory of mirages, 531
 Head, hydrostatic, 60, 139
 Hearing, 320
 Heat:
 conduction of, 253
 convection of, 250
 of fusion, 197
 mechanical equivalent of, 222
 quantity of, 185
 of solution, 241
 radiation of, 257
 specific, 186-190
 Hefner-Alteneck unit, 364
 Helion, 770
 Helium, nucleus of, 770
 Helmholtz's measurements:
 of pitch, 309
 of timbre, 309
 Henry, the, 684
 Hering's theory of color vision, 512
 Hertz's experiments, 718
 Hertzian waves, 521
 Heusler alloys, 552
 Hittorf's experiment, 728
 Horsepower, 58
 -hour, 59
 Hot-wire ammeter, 650
 Hue, 514
 Humidity, 211
 Hund's piezo-electric formula, 347
 Huygens' principle, 284
 applied to reflection, 286
 applied to refraction, 294
 applied to double refraction, 492
 Hydrostatic paradox, 120
 Hydrostatics, 116-128
 Hygrometers, 212
 Hypermetropia, 422
 Hysteresis, 668

I

Ice:
 effect of pressure on, 196, 197
 heat of fusion of, 198
 Iceland spar, 490
 Ideal gases, law of, 182
 Illumination, 364
 Images in plane mirror, 378
 inversion and perversion of, 379
 in lenses, 398, 401, 409
 size of, 388
 virtual, in lenses, 398, 399, 410
 virtual, in mirrors, 385, 388
 Impact, 112, 113
 Impulse, 34, 36
 Incandescent:
 gases, 519
 lamp, 614
 solids, 516
 Incidence, angle of, 287
 Inclination, magnetic, 554
 Inclined plane:
 as a machine, 72
 motion on, 51
 Indicator cards, 236
 Induced currents, 676-692
 e.m.f., 679
 magnitude of, 681
 quantity carried by, 681
 radioactivity, 797
 Inductance:
 mutual, 683
 self, 685
 Induction:
 coil, 687
 conditions of, 677
 electrostatic, 564
 lines of, 663
 motor armature, 711
 motor field, 709
 mutual, 682
 self, 685
 without motion, 679
 Inertia, 37
 moment of, 93, 94
 Infrared waves, 521
 Insulators of electricity, 560
 in electrostatic field, 572
 thermal, 262
 Interference, 290
 by Fresnel's biprism, 455
 by Newton's rings, 457, 458
 of light, 452-462
 of polarized light, 497
 of sound, 316
 by thin films, 457
 by two narrow apertures, 454
 Interferometer, Michelson's, 459
 used with telescope, 479
 Interval, in music, 328
 Intrinsic energy, 202, 225, 226
 Inverse square laws, 258
 Inversion of images, 379
 Inversion temperature:
 of gases, 231
 thermoelectric, 638
 Ionium, 800
 Ions:
 in solutions, 617
 migration of, 618
 Irreversible cycle, 235

Isobars, 800
 Isogonic lines, 555
 Isothermal, 131
 elasticity, 228
 processes, 226
 Isotopes, 770

J

Joly's steam calorimeter, 203
 Joule, the, 57
 Joule's determination of J , 223
 Joule's equivalent, 222
 Joule's experiments with gases, 229
 Joule-Thomson effect, 230, 231

K

Kaufmann's observations on electron
 mass, 734
 Kelvin scale, 175
 Kerr effect, 504
 Kilogram, 5
 Kilowatt-hour, 59
 Kinematics, 3
 Kinetic:
 energy, 60, 61, 64
 reaction, 38
 theory of gases, 178-183
 Kinetics, 3
 Kirchhoff's black-body ratio, 263, 264
 Kirchhoff's laws, 607
 Knot, nautical, 5
 Koenig's color-sensitivity curves, 511
 Kundt's tube, 357

L

Laevorotatory, 501
 Lambert, the, 366
 Lamp:
 arc, 614
 incandescent, 614
 Langley's radiation measurements, 612
 Langmuir's equation, 743
 Laplace's correction (sound velocity), 303
 Laplace's electromagnetic equation, 594, 596
 Laue photographs, 768
 Laurent's saccharimeter, 501
 Lawrence and Livingston:
 cyclotron, 785
 induced radioactivity, 798
 Leclanché cell, 632
 Lens, crystalline, 420
 Lenses, 397-415
 combinations of, 414
 focal length of, 402
 formula of, 405
 magnification by, 410
 optical center of, 408
 spherical aberration of, 411
 telephoto, 419
 thick, 413
 Lens's law, 676
 Lever, 70, 71
 Leyden jar, 579
 Light:
 cold, 529
 diffraction of, 487
 interference of, 452
 mechanical equivalent of, 528
 nature of, 363
 reflection of, 378

Light (*Cont.*):
 refraction of, 392
 sources of, 363, 511
 velocity of, 370-371
 wave length of, 43
 Limiting angle of repose, 68
 Limma, interval of, 31
 Linde's liquid-air machine, 220
 Linear:
 expansion, 163
 velocity, 9
 Liquefaction of gases, 220
 Liquid state, 214, 215
 Liquids, 116
 flow of, 138
 pressure in, 119
 Lissajous' figures, 335, 346
 Litter, 5
 Lodestone, 541
 Loudness, 322, 324
 Lumen, 364
 Luminescence, 527
 Luminosity of color, 515
 Luminous intensity, 363
 Lux, 364
 Lyman series, 526

M

Machines:
 efficiency of, 73
 simple, 69-74
 Magnetic:
 attraction, 541
 circuit, 671, 672
 dipoles, 548
 field, 543, 544
 flux, 662
 induction, 551
 lines of force, 546
 moment, 547
 rotation of polarized light, 503-505
 storms, 556
 Magnetism, 54-556
 dia-, 552
 ferro-, 551
 intensity of, 549
 of iron, 550
 para-, 552
 permanent, 551
 temperature effect on, 553
 terrestrial, 554
 Magnetization, curves of, 667
 Magnetomotive force, 670
 Magneto-optic phenomena, 503-505
 Magnetostrict ion, 549
 Magnets, 541
 unit pole of, 542
 Magnification:
 of compound microscope, 426
 of Galileo's telescope, 431
 of lenses, 410
 of mirrors, 388
 of reading glass, 425
 of reflecting telescope, 428
 of refracting telescope, 429
 of simple microscope, 424
 Major tone, 329
 Major triad, 330
 Manometers, 132
 Marconi's experiments with radio, 719
 Mariotte, law of, 130

Mass:
 defect, 789
 number, 787
 spectrographs, 771, 772
 Maximum and minimum thermometer, 161
 Maxwell:
 classification of colors, 514
 electromagnetic theory of light, 718
 theory of radiation pressure, 267
 Maxwell, the, 663
 Maxwell's "demon," 663, 674
 Mean free path, 116
 Mechanical:
 advantage, 71
 equivalent of heat, 222
 equivalent of light, 528
 Mechanics defined, 1
 Megabar, 118
 Melde's experiment, 339
 Melting:
 caused by pressure, 195
 point, 194
 Mendeleef's series, 767
 Meniscus, 148
 Mercury vapor lamp, 615
 Merit, figure of, 649
 Metacenter, 128
 Meter, the, 4
 Mho, the, 604
 Michelson and Morley's experiment, 461
 Michelson's determination of velocity of light, 373
 Michelson's interferometer, 459
 Microfarad, 658
 Micron, 435
 Microphone, 695
 Microscope:
 compound, 426
 simple, 424
 ultraviolet, 476
 Migration of ions, 618
 Miller:
 analysis of sound, 310
 experiments on "ether drift," 462
 Millicurie, 796
 Millikan's oil-drop experiment, 731
 Minimal surfaces, 151
 Minor tone, 329
 Minor triad, 330
 Mirage, 531
 Mirrors:
 concave, 383
 convex, 385
 parabolic, 389
 parallel, 381
 plane, 378-390
 rotating, 380
 two, 380
 Mixing colors, 507
 Mixtures:
 freezing, 244
 of vapors and gases, 209
 Modulus:
 of elasticity, 110
 of rigidity, 111
 shear, 111
 torsion, 111
 Young's, 110
 Molar solution, 624
 Mole, 624

Molecular:
 concentration, 624
 conductivity, 624
 forces, 145
 "hypothesis," 176
 range, 145
 weight, 181
Moment:
 of a force, 22
 of inertia, 93, 94
 of momentum, 97
Momenta, equilibrium of, 40-42
Momentum, 34
 compared to kinetic energy, 63
 conservation of, 37
 dimensions of, 36
 moment of, 97
Moseley's law, 778, 779
Motion, 8
 accelerated, 9
 circular, 79, 86
 Newton's laws of, 33
Motors:
 A.C. induction, 709
 D.C., 704-709
 efficiency of, 707
 regulation of, 708
 torque of, 707
Music, physics of, 328-335
Musical interval, 328
Mutual induction, 682
Myopia, 421

N

Negative:
 charge, 559
 crystal, 492
 glow, 726
 pole, 541
Neon lamp, 616
Nernst lamp, 611
Nernst's theory of voltaic cell, 629
Neutral equilibrium, 65
Neutral temperature (thermoelectric), 638
Neutron, 768
 discovery of, 784
Neutrons, production of, 787
Newton:
 law of cooling, 266
 laws of motion, 33, 34
 theory of sound propagation, 301
Newton's rings, 458
Nichols and Hull's radiation experiment, 267
Nichols and Tear's wave measurements, 521
Nicol's prism, 494
Nodes, 292
Nonconductors, 560
Normal solution, 624
Nucleus:
 composition of, 787
 of atom, 769

O

Objective, 426
Octave, 328
Ocular, 426
Oersted, the, 545
Oersted's experiment, 592

Ohm, 601
 international, 602
Ohm's law, 603
Onnes' experiments with superconduc-
 tivity, 612
Onnes' production of low temperatures,
 174
Opera glass, 430
Optic axis, 491
Optic center, 408
Orbits, atomic, 775
Organ pipes, 353
Orthogonal systems, 73, 574
Oscillation, center of, 104
Oscillations:
 electrical, 715-721
 frequency of, 716
 of triode, 746
Oscillograph, cathode ray, 736
Osmosis, 248
Osmotic pressure, 248
Overtones, 310

P

Packing effect, 788
Packing fraction, 788
Parallelogram of vectors, 11
Paramagnetic susceptibility, 664
Paramagnetism, 552
Parson's views on color vision, 514
Partials, of musical tones, 310
Pascal's principle, 118, 119
Paschen's law, 727
Paschen's series, 520
Peltier effect, 637
 cause of, 641
Pendulum:
 compound, 103
 conical, 84
 energy of, 90
 Foucault's, 106
 gridiron, 164
 simple, 89, 103
 torsion, 102
Percussion, center of, 103
Period of harmonic motion, 87, 89
Periodic series of elements, 767
Permalloy, 552, 668
Permeability, 663, 664, 667
Permeance, 672
Perrin's experiment, 668
Perrin's determination of Avogadro's
 number, 177
Perversion of images, 379
Pfeffer's osmotic membranes, 249
Phase:
 areas on p-v diagram, 217
 diagram, 216
 of wave motion, 277
Phonodeik, Miller's, 310
Phosphorescence, 524, 525
Photoelectric cell, 750, 752
Photoelectric phenomena, 748
Photographic camera, 417
Photographic lens rating, 418
Photometer:
 Bunsen's, 368
 Lummer-Brodhun's, 368
 Rumford's, 367

- Photometry, 367
 Photon, 518, 725
 Photosphere, 440
 Photronic cell, 751
 Physics, scope of, 1
 Piezo-electric oscillations, 346, 347
 Pigments, mixtures of, 507
 Pisa, leaning tower of, 65
 Pitch, 308
 measurement of, 344
 standards of, 331
 Planck's radiation formula, 518
 Plane, inclined, 72
 Plates:
 Chladni's, 342
 colors of thin, 499
 Poggendorf's potentiometer, 654
 Polarimeters, 502
 Polarization:
 by double refraction, 492
 by reflection, 487
 by scattering, 496
 electrolytic, 623
 of cells, 632
 rotatory, 500
 Polarized light, 486-505
 by Nicol's prism, 494
 by Polaroid, 496
 interference of, 497
 rotation of plane of, 500
 vector diagrams of, 493
 Polarizing angle, 489
 Polaroid, 496
 Poles:
 e.m.u., 542
 induced, 550
 north and south, 541
 Polonium, 791, 800
 Porous-plug experiment, 230
 Positive:
 charge, 559
 column, 726
 crystal, 492
 pole, 541
 Positron, 559, 763
 Post-office bridge, 656
 Potential:
 difference, e.m.u., 602, 603
 electrostatic, 567
 energy, 59
 gradient, 568
 gravitational, 60
 Potentiometer, Poggendorf's, 654
 Poundal, 35
 Power, dimensions and units of, 53
 Precession of gyroscope, 106
 Pressure, 21
 center of, 120
 change due to surface films, 149
 gradient, 139
 of atmosphere, 122
 of gas, 178
 Prévost's theory of exchanges, 260
 Primary colors, 510
 Principal planes, 413
 Principal points, 413
 Prism binocular, 431
 Projectiles, trajectory of, 50, 51
 Propagation of heat, 250
 Proton, 559, 768
 Pulleys, 73, 74
 Pump, lift, 123
 Pyrometer:
 resistance, 612
 thermojunction, 643
 Q
 Quadrant electrometer, 562
 Quality of sounds, 307, 309
 Quantity, e.m.u., 600
 Quantity of heat, 185
 Quantum, 518
 condition, Bohr's, 774
 Quartz:
 half-wave, 502
 oscillator, 346, 347
 Quincke's experiment, 145
 R
 Radial acceleration, 80
 Radian, 7
 Radiation:
 and temperature, 265
 mass equivalent of, 269
 pressure, 267, 268
 thermal, 257, 259
 Radioactive series, 800
 Radioactive transformations, 799
 Radioactivity, 790-803
 decay of, 795
 energy of, 796
 induced, 797
 measurement of, 798
 properties of, 791
 Radiometer, Crookes', 259
 Radius of gyration, 90
 Rainbow:
 colors of, 535
 form of, 533
 secondary, 535
 Raman effect, 527
 Ramsay's study of Brownian movements, 177
 Raoult's law of solutions, 243
 Ratio:
 of charge to mass, 731, 734
 of specific heats, 190, 357
 Rayleigh's criterion of resolution, 475
 Rays:
 alpha, beta, and gamma, 793, 794
 Becquerel, 790
 from beryllium, 784
 Reading glass, 425
 Réaumur scale, 160
 Recalescence, 553
 Reed pipes, 355
 vibrating, 351
 Reflection, 284
 angle of, 287
 change of phase in, 288, 289
 law of, 294
 of light, 378
 of sound, 312
 of thermal radiations, 261
 total, 394
 Reflectivity, 261
 Refraction, 284
 angle of, 293
 by prism, 394
 double, 489
 law of, 294
 of light, 392-396

- Refraction (*Cont.*)
 of sound, 314
 relative deviation of, 443
 Refrigeration, 219
 Regnault:
 measurements of sound velocity, 300
 measurements of thermal expansion, 166, 168, 172
 measurements of vapor pressure, 207
 Reluctance, 672
 Reluctivity, 672
 Remanence, 669
 Repose, limiting angle of, 68
 Resistance:
 e.m.u., 601
 internal, of cells, 626
 standard, 646
 temperature coefficient of, 605, 610
 Resistivity, 604
 table of, 695
 Resolution of vectors, 14
 Resolving power:
 calculation of, 477
 of lens, 475
 of slit, 475
 Resonance, 295, 296
 of sound, 351
 spectra, 450
 Restitution, coefficient of, 101, 102, 113
 Resultant, 12, 15
 of forces, 18
 Retentivity, 669
 Retina, 419
 Reverberation, time of, 325, 326
 Reversal of spectral lines, 441
 Reversible cycle, 231
 efficiency of, 235
 Reversing layer, 440
 Rheostats, 646
 Rigidity, modulus of, 111
 Ripples, 284
 Römer's determination of velocity of light, 370
 Röntgen rays, 754
 Rotation:
 and translation compared, 98
 of plane of polarization, 500
 problems of, 99
 vectors of, 104
 Rotational actions and reactions, 96, 104
 Rotatory power, 501
 Rotor, 702
 Rowland's experiment with rotating charge, 598
 Ruhmkorff coil, 688
 Rumford's experiment with heat of friction, 221
 Rumford's photometer, 367
 Rutherford and Soddy's theory of radio-activity, 795
 Rutherford's disintegration of atom, 782
 Rydberg constant, 520
- S
- Sabine's reverberation formula, 325
 Saccharimeter, Laurent, 501
 Saturated solution, 240
 vapor, 206
 measurements of, 207
 p-t curve of, 209
 table of, 208
 Saturation current in discharges, 737
 Saturation of color, 515
 Scalar quantities, 10
 Scale:
 diatonic, 329
 major, 329
 minor, 331
 tempered, 334
 Scleroscope, 114
 Screw, as a machine, 73
 Second, the, 4
 Seebeck effect, 637
 Self-induction, 685
 Semitones, 335
 Separator, centrifugal, 85
 Sextant, Hadley's, 382
 Shear modulus, 110
 Short-sight, 421
 Simple harmonic motion, 86
 acceleration of, 88
 velocity of, 88
 Simple microscope, 424
 Simple pendulum, 89, 90, 103
 Siphon, 121
 Size, apparent, 376
 Sky, color of, 537
 Slide-wire bridge, 656
 Slip of induction motor, 712
 Slug, the, 47
 Snell's law, 392
 Soap bubbles, 150
 Sodium vapor lamp, 616
 Solar constant, 259, 267
 Solar spectrum, 440-442
 Solenoid, 661
 Solid state, 214, 215
 Solute, 239
 Solution, heat of, 241
 Solution pressure, 629
 Solutions, 239
 boiling point of, 245
 freezing point of, 242
 saturated, 240
 Solvents, 239
 solids as, 240
 Sound, 298, 299
 analysis of, 309
 audibility of, 323, 324
 intensity of, 307
 interference of, 316
 pitch of, 308
 ranging, 305
 reflection of, 312
 refraction of, 314
 timbre of, 309
 Space charge, 742
 Space lattice, 759
 Spark:
 coil, 686
 discharge, 725
 discharge potentials, 729
 spectra, 519
 Specific gravity, 125
 Specific heat, 186-188
 of gases, 189, 190, 357
 quantum theory of, 192
 Specific inductive capacity, 580
 Specific resistance, 604
 Spectral series, 519
 Spectrometer, 437
 Spectroscope, direct vision, 446

INDEX

Spectrum, 435
 absorption, 439
 analysis, 438
 arc, 519
 band, 439
 continuous, 438
 emission, 438
 flame, 519
 line, 438
 mass, 771
 production of, 778
 resonance, 450
 solar, 440-442
 spark, 519
 Speed, 10
 Spherical aberration:
 of lenses, 411
 of mirrors, 389
 Stability of flotation, 127
 Stable equilibrium, 65
 Standard units, 4
 Standards of pitch, 331
 Stark effect, 460
 State:
 change of, 194
 equation of, 218
 Statics, 3, 17
 Stator, 702
 Stefan's radiation law, 266
 Steiner's theorem, 95
 Stereoscope, 375
 Ster-radian, 364
 Stokes' law of fluorescence, 524
 Storage cell, 634
 Strain, 109
 Stress, 109
 Strings, vibration of, 337
 Subdominant triad, 330
 Sublimation, 204
 Sundogs, 537
 Superconductivity, 611
 Supercooling, 199
 Superheating of liquids, 201
 Supersonic vibrations, 347
 Surface:
 color, 507
 films, 146
 tension, 147
 table of, 154
 Susceptibility, 663
 diamagnetic, 664
 ferromagnetic, 664
 paramagnetic, 664
 Synthesis of tones, 310

T

Telegraph, 666
 Telephone, 694
 Telephoto lens, 419
 Telescope:
 astronomical, 427, 428
 Galileo's, 430
 terrestrial, 429
 Temperature, 157-158
 coefficient of expansion, 163, 169
 coefficient of resistance, 605, 610
 of inversion, of gases, 231
 of sun, 266
 thermoelectric, 638

Tempered scale, 334
 Tension, surface, 147
 Terrestrial magnetism, 554
 Terrestrial telescope, 429
 Tesla coil, 721
 Thermal:
 capacity, 186
 conductivity, 253
 energy, 180
 expansion, 159, 163
 Thermionic current, 742
 Thermionic emission, 741-746
 Thermocouple meter, 644
 Thermodynamics, 225
 first law of, 225
 second law of, 234
 Thermoelectric curve, 641
 Thermoelectricity, 637-644
 Thermoelectromotive force, 638
 calculation of, 642
 table of, 643
 Thermometer:
 gas, 173
 maximum and minimum, 161
 mercurial, 159
 Thermometric properties, 158
 Thermometric scales, 159, 160
 Thermopiles, 643
 Thomson effect, 639, 640
 Thomson's (G.P.) electron diffraction
 Thorium:
 radioactivity of, 790
 series, 802
 Three-element tube, 743
 Thyatron, 751, 752
 Timbre, 307, 309
 Time, 4
 angle in s.h.m., 87
 Toepler-Holtz machine, 584
 Toepler's pump, 134
 Tone:
 complex, 309
 difference, 317
 major, 329
 minor, 329
 whole, 335
 Tonic triad, 330
 Toroid, 665
 Torque, 23
 of motors, 707
 Torricelli's theorem, 142
 Torricelli's vacuum, 122, 123
 Torsion modulus, 111
 moment of, 111
 Torsional pendulum, 102
 Torsional vibrations, 342
 Total reflection, 394
 Tractive force of magnet, 665
 Trajectory of projectiles, 50, 51
 Transformations, radioactive, 71
 Transformer, 689
 Transmission, thermal, 261
 Transmissivity, 261, 262
 Transport phenomenon (viscosity)
 Transposition of musical scales,
 Triads, 330
 Triode, as oscillator, 746
 as receiver, 744-746
 Triple point, 214
 Tungar rectifier, 743
 Tuning fork, 341

U

Ultraviolet waves, 521

Unit:

- charge, 563
- planes, 413
- pole, 542
- quantity, 600

Units:

- relations between e.s.u. and c.m.u., 691
- standard, 4
- c.g.s., 5

Unstable equilibrium, 65

Uranium:

- activity of, 790
- electronic structure of, 777
- radium series, 800, 801

V

Valence:

- chemical, 777
- of colors, 514

Valves, electrical, 719, 720

Van de Graaff generator, 585

Van der Waals' equation, 218

Vapor:

- pressure, 200
- saturated, 206-209

Vaporization, 199

- heat of, 201

Vectors, 10-15

- addition of, 10, 12, 13
- of rotation, 104
- reduction of, 14
- subtraction of, 11

Velocity, 5, 9

- angular, 79
- of effusion, 143
- of light, 369-374
 - by Bradley, 372
 - by Fizeau, 372
 - by Foucault, 373
 - by Michelson, 373, 374
 - by Römer, 370
- of liquid flow, 139, 140
- of liquid jet, 142
- of sound, 299-305
 - by Kundt's tube, 357
 - effect of temperature on, 304
 - in different media, 349

"Vena contracta," 143

Vibrations, 295, 296

- electromagnetic, 715-721, 746
- of gases, 350
- of jets, 350
- of membranes, 355
- of plates, 342
- of reeds, 351
- of rods, 340
- of strings, 337, 338

Virtual image, 385

Virtuality, 117

Visible universe, 374

Visible light

- in the ear, 375
- in the eye, 509-514

defects of, 421-424

Visual purple, 512

Volcanoes, 197

Voltaic cell, 629

Volta's discovery, 628

Voltmeter, 650

electrostatic, 653

Volume expansion, 165

W

Water:

- density of, 126, 168
- equivalent, 186
- specific heat of, 188
- waves, 281-283

Watt, 58

Wattmeter, 651

Wave mechanics, 779

Waves, 275

- diffraction of, 284
- electromagnetic, 717
- equation of, 281
- interference of, 290
- gravitational, 281, 282
- length of, 277
- longitudinal, 278
- reflection of, 284
- refraction of, 284, 293
- stationary, 291, 293
- surface tension, 283

Weber's theory of magnetism, 548

Wedge as a machine, 73

Wehnelt interrupter, 688

Weight, contrasted with mass, 46

Welding, electric, 614

Weston cell, 633

Wheatstone's bridge, 655

Wheel and axle, 71

Wien's displacement law, 517

Wilson cloud chamber, 762

Wind instruments, 356

Wireless telegraphy, 719

Work, 57

- dimensions of, 58
- function, 749

X

X-ray spectra, 756

X-ray tubes, 754

X-rays, 522, 754-762

- characteristic, 757
- fluorescent, 757
- nature of, 756
- secondary, 757

Y

Yield point, 112

Young-Helmholtz theory of color vision, 510

Young's modulus, 110

Z

Zeeman effect, 450

Zero, absolute, 174

Zone plates, 466

Condensed Table of Natural Trigonometric Functions

Angle	Sin	Cos	Tan	Angle	Angle	Sin	Cos	Tan	Angle
0° 0'	.0000	1.000	.0000	90° 0'	5° 0'	.0872	.9962	.0875	85° 0'
10'	.0029	1.000	.0029	50'	10'	.0901	.9959	.0904	50'
20'	.0058	1.000	.0058	40'	20'	.0930	.9957	.0934	40'
30'	.0087	1.000	.0087	30'	30'	.0959	.9954	.0963	30'
40'	.0116	.9999	.0116	20'	40'	.0987	.9951	.0992	20'
50'	.0145	.9999	.0146	10'	50'	.1016	.9948	.1022	10'
1° 0'	.0175	.9999	.0175	89° 0'	6° 0'	.1045	.9945	.1051	84° 0'
10'	.0204	.9998	.0204	50'	10'	.1074	.9942	.1080	50'
20'	.0233	.9997	.0233	40'	20'	.1103	.9939	.1110	40'
30'	.0262	.9997	.0262	30'	30'	.1132	.9936	.1139	30'
40'	.0291	.9996	.0291	20'	40'	.1161	.9932	.1169	20'
50'	.0320	.9995	.0320	10'	50'	.1190	.9929	.1111	10'
2° 0'	.0349	.9994	.0349	88° 0'	7° 0'	.1219	.9926	.1228	83° 0'
10'	.0378	.9993	.0378	50'	10'	.1248	.9922	.1257	50'
20'	.0407	.9992	.0408	40'	20'	.1276	.9918	.1287	40'
30'	.0436	.9991	.0437	30'	30'	.1305	.9914	.1317	30'
40'	.0465	.9989	.0466	20'	40'	.1334	.9911	.1346	20'
50'	.0494	.9988	.0495	10'	50'	.1363	.9907	.1376	10'
3° 0'	.0523	.9986	.0524	87° 0'	8° 0'	.1392	.9903	.1405	82° 0'
10'	.0552	.9985	.0553	50'	10'	.1421	.9899	.1435	50'
20'	.0581	.9983	.0582	40'	20'	.1449	.9894	.1465	40'
30'	.0611	.9981	.0612	30'	30'	.1478	.9890	.1495	30'
40'	.0640	.9980	.0641	20'	40'	.1507	.9886	.1524	20'
50'	.0669	.9978	.0670	10'	50'	.1536	.9881	.1554	10'
4° 0'	.0698	.9976	.0699	86° 0'	9° 0'	.1564	.9877	.1584	81° 0'
10'	.0727	.9974	.0729	50'	10'	.1593	.9872	.1614	50'
20'	.0756	.9971	.0758	40'	20'	.1622	.9868	.1641	40'
30'	.0785	.9969	.0787	30'	30'	.1651	.9863	.1668	30'
40'	.0814	.9967	.0816	20'	40'	.1679	.9858	.1695	20'
50'	.0843	.9964	.0846	10'	50'	.1708	.9853	.1721	10'
5° 0'	.0872	.9962	.0875	85° 0'	10° 0'	.1737	.9848	.1760	80° 0'
Angle	Cos	Sin	Cot	Angle	Angle	Cos	Sin	Cot	Angle

Angle	Sin	Cos	Tan	Angle	Angle	Sin	Cos	Tan	Angle
10° 0'	.1737	.9848	.1763	80° 0'	16° 0'	.2756	.9613	.2868	74° 0'
10'	.1765	.9843	.1793	50'	10'	.2784	.9605	.2899	50'
20'	.1794	.9838	.1823	40'	20'	.2812	.9596	.2931	40'
30'	.1822	.9833	.1853	30'	30'	.2840	.9588	.2962	30'
40'	.1851	.9827	.1884	20'	40'	.2868	.9580	.2994	20'
50'	.1880	.9822	.1914	10'	50'	.2896	.9572	.3026	10'
11° 0'	.1908	.9816	.1944	79° 0'	17° 0'	.2924	.9563	.3057	73° 0'
10'	.1937	.9811	.1974	50'	10'	.2952	.9555	.3089	50'
20'	.1965	.9805	.2004	40'	20'	.2979	.9546	.3121	40'
30'	.1994	.9799	.2035	30'	30'	.3007	.9537	.3153	30'
40'	.2022	.9793	.2065	20'	40'	.3035	.9528	.3185	20'
50'	.2051	.9788	.2095	10'	50'	.3063	.9520	.3217	10'
12° 0'	.2079	.9782	.2126	78° 0'	18° 0'	.3090	.9511	.3249	72° 0'
10'	.2108	.9775	.2156	50'	10'	.3118	.9502	.3281	50'
20'	.2136	.9769	.2186	40'	20'	.3145	.9492	.3314	40'
30'	.2164	.9763	.2217	30'	30'	.3173	.9483	.3346	30'
40'	.2193	.9757	.2248	20'	40'	.3201	.9474	.3378	20'
50'	.2221	.9750	.2278	10'	50'	.3228	.9465	.3411	10'
13° 0'	.2250	.9744	.2309	77° 0'	19° 0'	.3256	.9455	.3443	71° 0'
10'	.2278	.9737	.2339	50'	10'	.3283	.9446	.3476	50'
20'	.2306	.9730	.2370	40'	20'	.3311	.9436	.3509	40'
30'	.2335	.9724	.2401	30'	30'	.3338	.9426	.3541	30'
40'	.2363	.9717	.2432	20'	40'	.3366	.9417	.3574	20'
50'	.2391	.9710	.2462	10'	50'	.3393	.9407	.3607	10'
14° 0'	.2419	.9703	.2493	76° 0'	20° 0'	.3420	.9397	.3640	70° 0'
10'	.2447	.9696	.2524	50'	10'	.3448	.9387	.3673	50'
20'	.2476	.9689	.2555	40'	20'	.3475	.9377	.3706	40'
30'	.2504	.9682	.2586	30'	30'	.3502	.9367	.3739	30'
40'	.2532	.9674	.2617	20'	40'	.3529	.9357	.3772	20'
50'	.2560	.9667	.2648	10'	50'	.3557	.9346	.3805	10'
15° 0'	.2588	.9659	.2680	75° 0'	21° 0'	.3584	.9336	.3839	69° 0'
10'	.2616	.9652	.2711	50'	10'	.3611	.9325	.3872	50'
20'	.2644	.9644	.2742	40'	20'	.3638	.9315	.3906	40'
30'	.2672	.9636	.2773	30'	30'	.3665	.9304	.3939	30'
40'	.2700	.9629	.2805	20'	40'	.3692	.9294	.3973	20'
50'	.2728	.9621	.2836	10'	50'	.3719	.9283	.4007	10'
16° 0'	.2756	.9613	.2868	74° 0'	22° 0'	.3746	.9272	.4040	68° 0'
Angle	Cos	Sin	Cot	Angle	Angle	Cos	Sin	Cot	Angle

Angle	Sin	Cos	Tan	Angle
22° 0'	.3746	.9272	.4040	68° 0'
10'	.3773	.9261	.4074	50'
20'	.3800	.9250	.4108	40'
30'	.3827	.9239	.4142	30'
40'	.3854	.9228	.4176	20'
50'	.3881	.9216	.4211	10'
23° 0'	.3907	.9205	.4245	67° 0'
10'	.3934	.9194	.4279	50'
20'	.3961	.9182	.4314	40'
30'	.3988	.9171	.4348	30'
40'	.4014	.9159	.4383	20'
50'	.4041	.9147	.4418	10'
24° 0'	.4067	.9136	.4452	66° 0'
10'	.4094	.9124	.4487	50'
20'	.4120	.9112	.4522	40'
30'	.4147	.9100	.4557	30'
40'	.4173	.9088	.4592	20'
50'	.4200	.9075	.4628	10'
25° 0'	.4226	.9063	.4663	65° 0'
10'	.4253	.9051	.4699	50'
20'	.4279	.9038	.4734	40'
30'	.4305	.9026	.4770	30'
40'	.4331	.9013	.4806	20'
50'	.4358	.9001	.4841	10'
26° 0'	.4384	.8988	.4877	64° 0'
10'	.4410	.8975	.4913	50'
20'	.4436	.8962	.4950	40'
30'	.4462	.8949	.4986	30'
40'	.4488	.8936	.5022	20'
50'	.4514	.8923	.5059	10'
27° 0'	.4540	.8910	.5095	63° 0'
10'	.4566	.8897	.5132	50'
20'	.4592	.8884	.5169	40'
30'	.4618	.8870	.5206	30'
40'	.4643	.8857	.5243	20'
50'	.4669	.8843	.5280	10'
28° 0'	.4695	.8830	.5317	62° 0'
Angle	Cos	Sin	Cot	Angle

Angle	Sin	Cos	Tan	Angle
28° 0'	.4695	.8830	.5317	62° 0'
10'	.4720	.8816	.5355	50'
20'	.4746	.8802	.5392	40'
30'	.4772	.8788	.5430	30'
40'	.4797	.8774	.5467	20'
50'	.4823	.8760	.5505	10'
29° 0'	.4848	.8746	.5543	61° 0'
10'	.4874	.8732	.5581	50'
20'	.4899	.8718	.5619	40'
30'	.4924	.8704	.5658	30'
40'	.4950	.8689	.5696	20'
50'	.4975	.8675	.5735	10'
30° 0'	.5000	.8660	.5774	60° 0'
10'	.5025	.8646	.5812	50'
20'	.5050	.8631	.5851	40'
30'	.5075	.8616	.5891	30'
40'	.5100	.8602	.5930	20'
50'	.5125	.8587	.5969	10'
31° 0'	.5150	.8572	.6009	59° 0'
10'	.5175	.8557	.6048	50'
20'	.5200	.8542	.6088	40'
30'	.5225	.8526	.6128	30'
40'	.5250	.8511	.6168	20'
50'	.5275	.8496	.6208	10'
32° 0'	.5299	.8481	.6249	58° 0'
10'	.5324	.8465	.6289	50'
20'	.5348	.8450	.6330	40'
30'	.5373	.8434	.6371	30'
40'	.5398	.8418	.6412	20'
50'	.5422	.8403	.6453	10'
33° 0'	.5446	.8387	.6494	57° 0'
10'	.5471	.8371	.6536	50'
20'	.5495	.8355	.6577	40'
30'	.5519	.8339	.6619	30'
40'	.5544	.8323	.6661	20'
50'	.5568	.8307	.6703	10'
34° 0'	.5592	.8290	.6745	56° 0'
Angle	Cos	Sin	Cot	Angle

Angle	Sin	Cos	Tan	Angle	Angle	Sin	Cos	Tan	Angle
34° 0'	.5592	.8290	.6745	56° 0'	40° 0'	.6428	.7660	.8391	50° 0'
10'	.5616	.8274	.6788	50'	10'	.6450	.7642	.8441	50'
20'	.5640	.8258	.6830	40'	20'	.6472	.7623	.8491	40'
30'	.5664	.8241	.6873	30'	30'	.6495	.7604	.8541	30'
40'	.5688	.8225	.6916	20'	40'	.6517	.7585	.8591	20'
50'	.5712	.8208	.6959	10'	50'	.6539	.7566	.8642	10'
35° 0'	.5736	.8192	.7002	55° 0'	41° 0'	.6561	.7547	.8693	49° 0'
10'	.5760	.8175	.7046	50'	10'	.6583	.7528	.8744	50'
20'	.5783	.8158	.7089	40'	20'	.6604	.7509	.8796	40'
30'	.5807	.8141	.7133	30'	30'	.6626	.7490	.8847	30'
40'	.5831	.8124	.7177	20'	40'	.6648	.7470	.8899	20'
50'	.5854	.8107	.7221	10'	50'	.6670	.7451	.8952	10'
36° 0'	.5878	.8090	.7265	54° 0'	42° 0'	.6691	.7431	.9004	48° 0'
10'	.5901	.8073	.7310	50'	10'	.6713	.7412	.9057	50'
20'	.5925	.8056	.7355	40'	20'	.6734	.7392	.9110	40'
30'	.5948	.8039	.7400	30'	30'	.6756	.7373	.9163	30'
40'	.5972	.8021	.7445	20'	40'	.6777	.7353	.9217	20'
50'	.5995	.8004	.7490	10'	50'	.6799	.7333	.9271	10'
37° 0'	.6018	.7986	.7536	53° 0'	43° 0'	.6820	.7314	.9325	47° 0'
10'	.6041	.7969	.7581	50'	10'	.6841	.7294	.9380	50'
20'	.6065	.7951	.7627	40'	20'	.6862	.7274	.9433	40'
30'	.6088	.7934	.7673	30'	30'	.6884	.7254	.9490	30'
40'	.6111	.7916	.7720	20'	40'	.6905	.7234	.9545	20'
50'	.6134	.7898	.7766	10'	50'	.6926	.7214	.9601	10'
38° 0'	.6157	.7880	.7813	52° 0'	44° 0'	.6947	.7193	.9657	46° 0'
10'	.6180	.7862	.7860	50'	10'	.6968	.7173	.9713	50'
20'	.6202	.7844	.7907	40'	20'	.6988	.7153	.9770	40'
30'	.6225	.7826	.7954	30'	30'	.7009	.7133	.9827	30'
40'	.6248	.7808	.8002	20'	40'	.7030	.7112	.9884	20'
50'	.6271	.7790	.8050	10'	50'	.7051	.7092	.9942	10'
39° 0'	.6293	.7772	.8098	51° 0'	45° 0'	.7071	.7071	1.000	45° 0'
10'	.6316	.7753	.8146	50'					
20'	.6338	.7735	.8195	40'					
30'	.6361	.7716	.8243	30'					
40'	.6383	.7698	.8292	20'					
50'	.6406	.7679	.8342	10'					
40° 0'	.6428	.7660	.8391	50° 0'					
Angle	Cos	Sin	Cot	Angle	Angle	Cos	Sin	Cot	Angle

CENTRAL LIBRARY

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE

PILANI (Rajasthan)

Call No.

530

P41C.

Acc. No.

11251

DATE OF RETURN

--	--	--	--

